IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-4, Special Issue-5, Jul 2016 E-ISSN: 2347-2693

Review on Big Data and Associated Technologies

Rimmy Yadav[1*], Anil Sharma[2]

[1]School of Computer Application, Lovely Professional University, India
[2]School of Computer Application, Lovely Professional University, India

Available online at: www.ijcseonline.org

Abstract- Global business, business people, government are generating and consuming vast amount of data very frequently. So in the world of digital computing, it's become very challenging to deal with this variety and velocity of the data. To overcome these challenges, Big Data is playing an important role by providing, capturing, managing and analyzing these big data sets. Software and technologies of big data facilitates the increase in organizational growth. To know about complete big data, focused has made on the emerging technologies i.e. cloud computing, Internet of Things (IoT) and Hadoop that are closely related to the big data. For each related technology, a description and key features have been highlighted. Finally the authors examined the relationship and benefits among big data and its associate technologies.

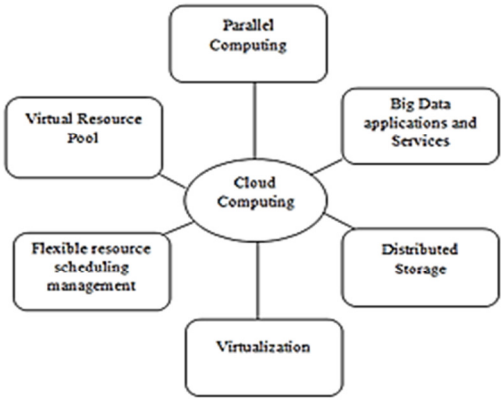
Keywords-Big Data, Cloud Computing, Hadoop framework. Internet of Things (IOT).

I. BIG DATA

Big data refers to high volume of heterogeneous datasets which cannot be easily managed by the traditional database management tools. The emergence of big data significantly increased the managerial, tactical and operational capabilities of an organization in terms of data processing, information retrieval, privacy and security, and most importantly decision making. This all is possible through advanced big data tools and techniques. The continuous adoption of Big Data Systems (BDS) are rapidly providing various advantages such as (a) storing vast amount of data (in terms of Petabyte (10^{15}), Exabyte (10^{18}), and Zettabyte (10^{21}), (b) structure (text based data) and unstructured data (images, audio, video etc.), (c) data can be saved over a long period of time, (d) helps decision makers to seek unforeseen data predictions throughout the datasets to make better judgment to grow in the market place [1]. The big data has the capability to hold vast amount of semi- structured and unstructured data in the databases. Data could be of any type such as internet data (e.g. web pages); Internet of Things (IoT) based data, enterprise data (such as inventory data, financial data, and sales data), etc. Big data employs high speed communication network such as orthogonal frequency- division multiplexing (OFDM), data centers (DC) which are clusters of varying number of servers, data pre- processing techniques such as redundancy elimination, data cleaning etc., [2]. In addition to this big data also employ various business statistical techniques such as projecting analysis, machine learning, mining of data and various statistical methods such as R programming model, Microsoft Excel, Pentaho/ weka to see unforeseen future prediction from vast amount of databases which then help the decision makers to implement beneficial strategy to promote their business [3][4][5].

II. CLOUD COMPUTING

The cloud computing has change the entire facet of the information technology (IT) which helps the leading organizations whether it is large scale or medium scale [6][7][8]. Traditional information processing systems are now adapting the growing and emerging technologies of the cloud computing. The cloud computing is the next generation of distributed computing. With the help of cloud computing the resources can used at any location. With the help of cloud computing, client can acquire any kind of needed services with a subscription amount. Such service may be hardware, software, etc. Cloud computing has the capability to deliver the resources/ services either local



cloud users or to users of geographical distributed locations. The components of cloud computing paradigm are illustrated in Fig.1.

A. *Cloud Service Model*

Cloud computing provide offers various valuable services to its intended users, it could be software or hardware or anything. These services are provided by means of virtualization. Cloud service provider can offer vast and varying number of data storage devices and computational resources. The service layer architecture of cloud computing paradigm is depicted in Fig. 2.

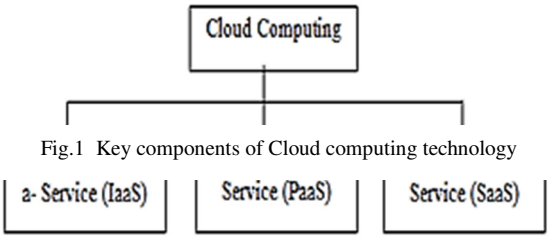


Fig.1 Key components of Cloud computing technology

Fig.2 Service layers of Cloud Computing

Basically cloud provides three computing service model. These are IaaS, PaaS and SaaS.IaaS stands for Infrastructure- as- a- Service, PaaS for Platform- as- a-Service and SaaS is term which stands for Software- as- a-Service.IaaS offers various services such as storage devices, processors and other computational devices. These services can be provided by means of virtualization. In PaaS, the customer creates the software using various software and other utilities provided by the service provider.A client user can also easily manage the software deployments and other configuration setting [8]. In SaaS, the software's are provided to the customers according to his needs. It is also known as "on demand service model".

A. *Important features of cloud computing*

- Services such as email, applications, network or servers can be offered to cloud clients with minimal subscription charges. Major cloud service providers are Amazon EC2 and Microsoft.[9]
- Network access to sundry areas: Cloud computing has the capability to deliver its services to any location of any region. It can be a sundry area or smart citythrough the mobile devices, laptops and PDAs [10][11].
- Resource Sharing: Virtualization features of cloud computing enables the sharing of storage devices and computation resources. [12].
- Elasticity: The facilities offered by the cloud service provider will be scaled to a large distributed location depending upon the mode of subscription.[13]

B. *Cloud computing with big data*

The development of big data hurries the progress of cloud computing environment. Cloud computing provides various benefits to big data. Table 1, shows the advantages of cloud computing withbig data platform are:

TABLE I
Benefits of cloud computing to big data

Ref. No.	Benefits of cloud computing to big data
[13]	Cloud computing provides large- scale physical resources such as storage and computing resources to big data platform to process vast amount of data.
[14]	Cloud computing platform have inbuilt resource scheduling and work- load management capabilities which then helps the big data systems to manage large data sets.
[14]	Reduce infrastructure installation cost and administration cost.
[15]	Business Intelligence-as- s- service (BIaaS) layer of cloud computing delivers various business analytic software such as R programming tool, weka for clustering and classification of large datasets which helps decision makers to make profitable decisions.
[16]	Remote Data Auditing (RDA) techniques of cloud computing efficiently protect the users' private information stored in the data sets.
[17]	Provides Reliable Computational frameworks such as Hbase, Sailfish i.e. similar to the Hadoop map- reduce programming model, helps to save vast variety of unstructured and semi-structured datasets.

Cloud computing not only provides computational and processing for big data, but also is a service mode. The improvements of cloud computing also support the progress of big data, both of which supplement each other.

III. INTERNET OF THINGS

The concept of IoT relies on the use of numerous heterogeneous virtual machines, which are connected to the Internet.This platform consists of already existing and emerging Internet developments. The basic idea of the IoT is to connect the numerous devices such as sensors, BCR (Bar Code Reader) and mobile phones, to realize information exchange and collaboratively complete the given task efficiently. For example: sensors can be used to acquire and generate weather forecasting report, sensors embedded within a mobile device can communicate with Global Positioning System (GPS) to identify the location [17]. Sensors with Radio wave frequency enabled devices can be used for underwater communication to identify

drowned objects.Following are the major Key components of IoT.

A. Key Components of IoT

- Sensor based technology: Vast amount of information e.g. humidity, pressure, weather forecasting report can be transfer to any other location.
- IoT gateways: It is used to connect the internal network of sensors with the external internet. The collected data is then transmitted to the www or internet infrastructure.
- Cloud/ Server infrastructure &big data: The data transmitted via the gateway is stored and processed securely within the cloud infrastructure using big data analytical engine. Because cloud computing include large computing resources and storage systems. Analytical engine is then used to seek the important or relevant information from stored IoT datasets
- End- user mobile applications: These mobile applications will help end users to control and monitor their devices from remote locations [19].

Consider the Fig. 3 to illustrate the working of IoT technology.

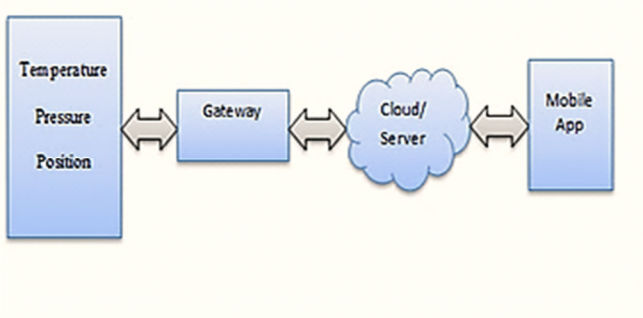


Fig. 3 IoT technology paradigm

B.Benefits of IoT

- With the help of sensors and wireless technology, health care equipment's can easily transfer the patient data to the authorized doctor or nurse who have permission to access the mobile application.
- Information such as energy consumption within different departments of an organization can be gathered using IoT based technology.
- IoT technology enables the development of smart cities and urban planning using big data analytics. [20]
- IoT is widely adopting in the field of agriculture which helps the farmers to see the fertility level of soil, humidity, temperature, pressure etc.

C. Relationship of IoT with Big data.

TABLE II

Benefits of IoT with big data

Ref. No	Benefits of IoT with Big data
[17]	IoT devices generates enormous amounts of data and transmits it to the business intelligence and business analytics tools for humans to make beneficial decisions such as to increase customer satisfaction, market trends etc.
[18][19]	IoT plays a vital role in organizations where employees can get their tasks on IoT enable devices.
[20][21]	Using big data analytics with IoT technology paradigm, consumer can manage their basic needs such as water usage and water sustainability.
[21]	IoT technology paradigm can make a use of Hadoop framework to increase its security, scalability of IoT datasets.
[22]	Big data and Hadoop framework can be used for outlier detection and duplicate content elimination in IoT datasets.

IoT generates vast amount of image based data. adoption of big data platform and its services not only the provide the facility of storing the vast amount of IoT datasets but also provides other services such as privacy and security, high processing and scalability capabilities to IoT datasets.

IV. HADOOP

In this digital world of computing each and every field is generating vast amount of data either it will be in the form of text or audio and video based content.Processing and analyzing such kind of these large quantity of datasets and to mine meaningful information is a challenging task. This will require a large scale data intensive applications and storage clusters along with requirements such as fault tolerance, scalability, parallel processing, load balancing and high availability. Hadoop is the best solution for the above said requirements [23].

Hadoop is open source implementation of map-reduceprocessing model. Hadoop is a set of software framework along with ongoing projects such as Pig, Hive and Hbase for consistent, scalable, similar and distributed computing[24]. Hadoop is framework of software tools which enables the applications to run smoothly on the big data platform. Hadoop Distributed File System (HDFS) and Map- Reduce processing model are two major components. Hadoop runs on low commodity hardware devices such as Linux operating system based embedded devices. HDFS is a highly accessible and fault tolerantin nature and is able to store file which is big in size across multiple nodes in a cluster. E.g. HDFS automatically creates the 3 replicated copies (by default) of original data in distributed storage cluster. On the other hand Map- Reduce is programming model which is used to process and analyze large amount of data sets. Map- reduce is similar to the Google File System (GFS) [25] which can be used to manage and process geographical parallel- distributed datasets.Hadoop HDFS can save structured, semi- structured and unstructured datasets whereas Map- Reduce has the ability to perform computations over these datasets.

A. Benefits of Hadoop

- Scalability:Hadoop frameworkpermits the development and shrinkage of hardware infrastructure without changing the format of data.
- High cost proficiency: Hadoop applies large- scale parallel computing to commodity hardware devices, which greatly reduce the storage cost [25].
- Flexibility: it may manage different kinds of datasets from various data sources [26].
- Fault- tolerance: Hadoop has inbuilt fault tolerance functionality. Task of one node can be easily replicate to other nodes by the HDFS. So in case of failure of task it can be resumed from point where it has stop working due to node failure.

B. Benefits of Hadoop and Big Data

The following table shows the tremendous advantages of Hadoop with big data.

TABLE III
Benefits of Hadoop with big data

Sr. No.	Benefits of Hadoop with big data.
1	Hadoop framework can be used to manage and analyze the big data social networking applications such as Facebook, twitter, etc. [26]
2	Hadoop ongoing project named as Hbase not only save the unstructured data but also enables the storage of structured datasets.
3	With the help of Pig, one of Hadoop Project, enables the programmers to write Map- reduce based programs for storing big data.
4	Hadoop framework with various business analytic tools such as R programming model, weka, etc. can be used to classify the relevant and irrelevant data from big data sets. [27]
5	Map- Reduce programming model of Hadoop enables the programmers to scale the data mining algorithms on large data sets.[28]

Hadoop is acts as business analytic for big data platform. It has the ability to deal with any kind of failure while processing with big data sets.

V. CHALLENGES OF BIG DATA

The following shows the challenges that are arising in big data environment.

TABLE IV
Challenges of big data platform

Sr. No.	Challenges in Big Data
1	Security issues are the major concern in big data. One of the important security issues on the input part of the big data is to make sure that the sensors, log files, web crawlers etc. will not be compromised by attacks.
2	Security problem on the communication between the big data and other external system is also major concern.
3	Dealing with security issues on the analysis part of big data is also a challenging task.
4	To identify the Advanced Persistent (APTs) security issues through the disparate systems is a crucial task.
5	Leakage of private information by the data analytics techniques to other people after the big data analysis process is a challenging task.
6	Different convergence speeds of the same data mining algorithms leading to the problem of synchronization.
7	How to mitigate the impact of noise, outliers, incomplete and inconsistent data in big data storage is becoming an open issue.
8	Manually cleaning data in big data is considered as the main challenge in the arena of big data due to the increasing volume, velocity and variety of data.
9	Challenge is to develop a filtering mechanism to keep the useful information in big data.
10	Deciding the location where to store the big data is another challenging task.
11	To find the root cause failure of distributed compute nodes, databases middleware is an extremely laborious process.

VI. CONCLUSION

Rapid development in big dataimproves the status of cloud computing, IoT and Hadoop framework.Cloud computing provides various benefits such as storage devices which will save huge amount of big data with minimum cost. Issues such as (1)Migrating big data applications to cloud computing platform is a challenging task because user's private information will be exposed and renting the cloud infrastructure is very costly.(2) Failure of virtual node will slow down the performance of big data application in cloud environment.In [29], authors proposed an Ant Colony Optimization technique to provide shortest path during communication link failure. Such kind of techniques will be implemented in big data environment to increase the availability of the resources. Replication of task will be better option for big data so increases availability of resources [30].There is a stringent requirement of privacy and security mechanism which reduces the chance of information leakage while migration of big data to the cloud.Internet- of- things (IoT) technology paradigm do not have data preprocessing tools to eliminate redundancy from IoT datasets. IoT technology should employ data processing tools of big data for data cleansing and duplicate content elimination.On the other hand, Hadoop framework also confronted with many failures such as (i) Hadoop framework is not well comfortable with cloud infrastructure because Hadoop requires homogeneous commodity hardware whereas cloud provides heterogeneous types of computing resources.

For developing mechanism either for cloud based technology or for IoT and Hadoop based computational framework, the researcher, data analyst and cloud developers should adopt the intelligence of big data and

business analytics so that future of IT can be easily predict for beneficial purposes.

REFERENCES

[1] X. Zhang, C. Liu, S. Nepal, C. Yang and J. Chen, "Privacy preservation over big data in cloud systems."*Security, Privacy and Trust in Cloud Systems*.Springer Berlin Heidelberg, 2014, pp.239-257.

[2] H. Wang, X. Qin, X. Zhou, F. Li,Z. Qin,Q. Zhu andS. Wang, "Efficient query processing framework for big data warehouse: an almost join-free approach." *Frontiers of Computer Science* 9.2, 2015, pp.224-236.

[3] J. Merino, I. Caballero, B. Rivas, M. Serrano and M. Piattini,"A Data Quality in Use model for Big Data."*Future Generation Computer Systems*, 2015.

[4]E. Elsebakhi, F. Lee,E. Schendel, A. Haque, N. Kathireason,T. Pathare,N. Syed and R. Al-Ali, "Large-scale machine learning based on functional networks for biomedical big data with high performance computing platforms." *Journal of Computational Science* 11, 2015,pp. 69-81.

[5]C.H. Chao, "The Framework of Information Processing Network for Supply Chain Innovation in Big Data Era." *The 3rd International Workshop on Intelligent Data Analysis and Management*.Springer Netherlands, 2013, pp.77-85.

[6] E.N. Alkhanak,S. P. Lee,R. Rezaei andR.M. Parizi, "Cost optimization approaches for scientific workflow scheduling in cloud and grid computing: A review, classifications, and open issues." *Journal of Systems and Software* 113, 2016, pp.1-26

[7] F. Hao,X. Yi and E.Bertino, "Editorial of special issue on security and privacy in cloud computing."*Journal of Information Security and Applications* 27, 2016, pp. 1-2.

[8] B. Snyder, J. Ringenberg, R. Green, V.DevabhaktuniandM. Alam, "Evaluation and design of highly reliable and highly utilized cloud computing systems." *Journal of Cloud Computing: Advances, Systems and Applications* 4.1, 2015, pp.11.

[9] A. Iosup, S. Ostermann, M. N. Yigitbasi, R. Prodan, T. Fahringer and D. H. Epema, "Performance analysis of cloud computing services for many-tasks scientific computing." *Parallel and Distributed Systems, IEEE Transactions on* 22.6, 2011, pp.931-945.

[10]S. Distefano, G. Merlino and A. Puliafito, "A utility paradigm for IoT: The sensing Cloud." *Pervasive and mobile computing* 20, 2015, pp.127-144.

[11] A. U. R. Khan, M. Othman, F. Xia and A. N. Khan, "Context-Aware Mobile Cloud Computing and Its Challenges." *Cloud Computing, IEEE* 2.3, 2015, pp.42-49.

[12] M. N. Sadiku, S. M. Musa andO. D. Momoh, "Cloud computing: Opportunities and challenges." *Potentials, IEEE* 33.1, 2014, pp.34-36.

[13]Q. Zhang, C.Lu and B. Raouf, "Cloud computing: state-of-the-art and research challenges." *Journal of internet services and applications* 1.1 (2010), pp.7-18.

[14]M. Alrokayan,A. Vahid Dastjerdi and R. Buyya,"Sla-aware provisioning and scheduling of cloud resources for big data analytics."*Cloud Computing in Emerging Markets (CCEM), 2014 IEEE International Conference on*.IEEE, 2014, pp.1- 8.

[15] V. Chang, "The business intelligence as a service in the cloud." *Future Generation Computer Systems* 37, 2014, pp.512-534.

[16]M. Sookhak, A. Gani, M. K. Khan and R. Buyya, "Dynamic remote data auditing for securing big data storage in cloud computing."*Information Sciences*, 2015.

[17]M. Fazioand A. Puliafito, "Cloud4sens: a cloud-based architecture for sensor controlling and monitoring." *Communications Magazine, IEEE* 53.3, 2015, pp.41-47.

[18] I. Lee and K. Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises." *Business Horizons*58.4 , 2015, pp.431-440.

[19] M. Wei, S. H. Hong and M. Alam,"An IoT-based energy-management platform for industrial facilities."*Applied Energy* 164, 2016, pp.607-619.

[20] M. M. Rathore,A. Paul, A.Ahmad and S. Rho, "Urban planning and building smart cities based on the internet of things using big data analytics."*Computer Networks*, 2016.

[21] S. K. Sowe, T. Kimata,M. Dong and K. Zettsu,"Managing heterogeneous sensor data on a big data platform: IoT services for data-intensive science." *Computer Software and Applications Conference Workshops (COMPSACW), 2014 IEEE 38th International*.IEEE, 2014, pp.295- 300.

[22] A. M. Souza and J. R. Amazonas, "An Outlier Detect Algorithm using Big Data Processing and Internet of Things Architecture." *Procedia Computer Science* 52, 2015, pp.1010-1015.

