ISSN No. 0976-5697

Volume 8, No. 2, March 2017 (Special Issue)

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Big Data on Cloud: A Review

Parul Agarwal
Department of CSE, SEST
Jamia Hamdard, New Delhi, India
parul.pragna4@gmail.com

Siddhartha Sankar Biswas*
Department of CSE, SEST
Jamia Hamdard, New Delhi, India
ssbiswas1984@gmail.com

Abstract: "Big data", data which is too big and requires highly scalable tools, platforms, computing techniques in today's scenario calls for best known and parallel facility – "Cloud Computing". Both have emerged as an interrelated area of research. In this paper, the emerging issues in the research of Big data are solicited in context of Cloud are explored theoretically. It also gives a holistic view of research challenges associated with both. The prima focus of the paper would be parameters like analysis, data integrity, scalability and legal and privacy issues.

Keywords: Big Data,Volume, Velocity, Variety,Scalability

I. INTRODUCTION

We started with organizations maintaining just "data". Data which was being stored and used. Then came the era of this data not just being queried to get answers to questions to the one where the data stored was mined for knowledge. This was because automated data collection tools had led to large amount of data being stored in databases and other information repositories.

Mining was a process of identifying inherent interesting patterns from data. This knowledge was then incorporated for Business Intelligence. This model changed. Earlier, there were some as producers of data and others as consumers of data. But today, all of us are producing data

And all are consuming it as well. This process of production and consumption of data is because of the factors- Social networking sites, Web Data, Sensor data, increased use of scientific instruments, explosion of mobile technology and many others.

This data which is too big to handle is the "Big Data". Data, data which is continuously evolving, changing, is dynamic in nature, too large to store and handle. The hindrance in advancement lies in the fact associated with big data is its storage and processing. But the best solution is "Cloud Computing". It's a platform which lets you perform large scale and complex computing. Cloud computing has its own advantages in terms of minimized maintenance cost, efficient user access, reduced cost for automation. Thus, it proves to be a scalable, fault tolerant and reliable environment for big data systems to perform[4]. The need of the hour is to harness the advantages they provide if one is migrated onto another. In this paper we describe both Big data systems and cloud computing and provide a comprehensive review of how they work in sync with each other. We particularly touch upon issues like security, privacy, heterogeneity and others.

Section 2 describes Big Data, its definition, attributes, and a detailed description.

Section 3 provides an insight into the Cloud Computing concept, while focusing on its characteristics and its tools.

Section 4 discusses how cloud computing and big data are related and dependent on each other for their existence and significance.

Section 5 discusses the research challenges related to both and discusses aspects like scalability, security, privacy and heterogeneity.

Section 6 provides a conclusion of the study.

II. OVERVIEW TO BIG DATA

Big Data [1] is data which is massive and difficult to store, manage and process.Big data has been defined in various means[2-3]owing to its origin in reality. Big Data is characterized by an initial 3 V'S Model [6]and was later attributed by a fourth parameter thus making it as 4 V'S Model. 3V's Model, is the one where the V's denote Volume, Velocity and variety[5] .

Thus big data is not only being concerned with its storage but also with analysis, its processing and knowledge extraction [4]. Gradually from 4 V'S Model, where veracity was added we have shifted to the 5 V'S Model characterized by Volume, Variety, Velocity, Value and Veracity[7].

Volume: Volume refers to the huge amount of data which is generated every second[8]. This amount of data is measured in Pettabytes, Zettabytes and even Brontobytes. The source of such huge data is data coming from Social Networking sites in form of posts or tweets, emails, photos, sensor data, and videos that we produce and share. This data surely concerns scalability, performance, bandwidth and its availability.

Velocity: Velocity refers to therate at which this data is being generated and thus processed. The use of digital devices has led to an unprecedented rate of data creation.It thus concerns the different rates at which data enters and exits the system and provides an abstract ion level which can store it independent of the incoming or the outgoing data. Systems should be capable of processing data even with variable velocity [10].

Variety: Variety refers to the different types of data available to any organization [9,10]. This data is unstructured or semi-structured. But with the help of big data technologies, organization can easily handle

this data and make it in a format which can be processed. Thus variety is not only concerned with different types but also with its different uses and analysis [11].

Veracity: Veracity refers to the accuracy, correctness or trustworthiness of the data. Accuracy of data is very important for a correct analysis of the data to be done efficiently.

Vendors / Properties	Google	Microsoft	Amazon	Cloudera
Big Data Storage	Google Cloud Services	Azure	S3	N/A
MapReduce	AppEngine	Hadoop on Azure	Elastic MapReduce (Hadoop)	MapReduce YARN
Big Data Analytics	BigQuery	Hadoop on Azure	Elastic MapReduce (Hadoop)	Elastic MapReduce (Hadoop)
Relational Database	Cloud SQL	SQL Azure	Mysql Or Oracle	MySQL, Oracle, PostgreSQL
NoSQL Database	AppEngine Datastore	Table storage	Dynamodb	Apache Accumulo
Streaming Processing	Search API	Streaminsight	Nothing Prepackaged	Apache Spark
Machine Learning	Prediction API	HadoopþMahout	Hadoopþmahout	HadooppOryx
Data Import	Network	Network	Network	Network
Data Sources	A few sample datasets	Windows Azure marketplace	Public Datasets	Public Datasets
Availability	Some services in private beta	Some services in private beta	Public Production	Industries

Value: With advent of big data, its worthiness or importance has increased multi fold because it is being used to make intelligent decisions and derive knowledge out of knowledge.

Hadoop : The framework

Hadoop is a java based programming framework which is freely available and supports the processing of huge data in a distributed environment. It is used for processing the unstructured data. Hadoop can be used for processing large amount of data over several clusters of servers. Also, applications can be executed on systems with several hundreds of nodes which involves terabytes of data. Thus, it lowers the risk of failure even if nodes fail enabling a fault tolerant and a scalable environment. HDFS[12], the file system spans all nodes in a Hadoop cluster for data storage and improves reliability.

Challenges of Big Data

Big data is confronted with several challenges[13]. Efficient research efforts are required that deal with its storage, analysis and display[14]. Major challenges include the hardware and the software requirements[15], heterogeneous source of data[16], search of real time big data[17], handling redundant data to name a few.

III. Cloud Computing

A fault tolerant computing environment that supports increased capability of storage and processing. We all use cloud in some form or the other. Most common being the using applications such as iCloud, Gmail, Dropbox to name a few which have become prevalent today. The reason of increased use of cloud computing is increased use of wireless networks and mobile applications. So it's the need of organizations, individuals which has led to the need to store and process such huge data so as to adopt cloud computing[18].Cloud is associated with IaaS(Infrastructure as a Service), PaaS(Platform as a service) and SaaS(Software as a Service).

Big Data On Cloud

Big Data due to its challenges and its features requires a scalable and fault tolerant environment. And the solution is Cloud Computing. So an efficient amalgamation of both is needed so as to harness the advantages of both. Cloud Comuting offers the solution through hardware virtualization.

IV. Research Challenges

In the following sub sections we identify the issues related to big data and answer the solution in form of a cloud.

Scalability

The data is growing and is being generated as Perrabytes of data. How do I Store it? Where do I keep the data? What algorithms will be used for processing it? Will any Data Mining technique be able to handle such huge data? Several scalable techniques are being used by organizations such as Microsoft. The transfer of data onto the cloud is a slow process and we need a proper system that does it at a considerable speed especially when the data is dynamic in nature and huge. Data rebalance algorithms exist and are based on load equalization and histogram build up.[19]. Scalability exists at the three levels in the cloud stack. At the Platform level there is: horizontal and vertical scalability.

Security and Access Control

Security is an aspect that arises as a problem from inside an organization or when an individual uses a cloud to upload "its own data". When a Client uploads

a data and pays too for the service, so who is responsible for access to the data, permissions to use the data, the location of the data, its loss, authority to use the data being stored on clusters, The right of the cloud service provider to use the client's personal data and many others.One of the major solution was encrypting the data. But querying this data became a problem. Considerable answers have been provided but many still need to be answered and researched.[20-22]

Privacy and Integrity Issues

The data being generated might be too personal for an individual or an organization. This big data might be collected from Facebook accounts, WhatsApp applications each of these being more personal as compared to other applications. In addition to this online data, several data maybe pertaining to health records purchases etc. these might lead to,identification issues, profiling, loss of control, location whereabouts of a person related to purchases in supermarkets and many more.

Thus anonymization of this data or its encryption come as solutions to this issue. Privacy approaches can be dealt with user consent over its usage or sharing on the globe. Several privacy and protection laws exist for this which are a part of regulatory framework.

V. CONCLUSION

With this paper we have touched upon how big data and cloud provide solutions to each other. Though Cloud has proved to be a solution of various challenges of big data but still many challenges are facing both. These have to be identified and researched to get maximum benefits of both.

V. REFERENCES

[1] http://strata.oreilly.com/2012/01/what-is-big-data.html

[2] http://blog.softwareinsider.org/2012/02/27/mondays-musings-beyond-the-three-vs-of-big-data-viscosity-and-virality/

[3] O. R. Team (2011) Big data now: Current Perspectives from OReilly Radar. OReilly Media.

[4] I.A.T. Hashem, et al., The rise of "big data" on cloud computing: Review and open research issues, Information Systems(2014), http://dx.doi.org/10.1016/j.is.2014.07.006

[5] http://dashburst.com/infographic/big-data-volume-variety-velocity

[6] Chen, M., Mao, S., Liu, Y. "Big Data: A Survey" published online in Springer Science + business media, New York, 2014.

[7] Sakr, S. & Gaber, M.M., 2014. Large Scale and big data: Processing and Management Auerbach, ed.,

[8] Ammu, Nrusimham, and MohdIrfanuddin. "Big Data Challenges."International Journal of Advanced Trends in Computer Science and Engineering, 2:1,613

[9] Sagiroglu, Seref, and DuyguSinanc, 2013. "Big data: A review." International Conference on Collaboration Technologies and Systems (CTS), pp 42-47.

[10] Mayer-Schönberger, Viktor, and Kenneth Cukier. Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin, 2013.

[11] EMC:Data Science and Big data Analytics, 2012.: In: EMC Education Services, pp 1-508.

[12] K, Chitharanjan, and Kala Karun A. "A review on hadoop HDFSinfrastructure extensions.". JeJu Island: 2013, pp. 132-137, 11-12. Apr. 2013.

[13] Jagadish, H. V., et al. Univ. of Michigan (Coordinator)], "Challenges and Opportunities with Big Data", 2012.

[14] Shilpa, kaur, M. "Big data and Methodology - A review" in International Jouranal of Advanced Research in Computer Science and Software Engineering, Vol 3. 2013.

[15] Bakshi, K., "Consideration for big data : Architecture and Approach" Aerospace conference, 2012 IEEE (3-10 May'2013).

[16] Grobelnik M (2012) Big data tutoriual, http://videolectures.net/eswc2012grobelnikbigdata.

[17] Vaswani, G., Bhatia, A. "A Real Time approach with Big data - A Review", in International Jouranal of Advanced Research in Computer Science and Software Engineering, Vol 3.Issue 9, 2013.

[18] L. Huan, Big data drives cloud adoption in enterprise, IEEE Internet Comput. 17 (2013) 68–71.

[19] Mahesh, A. et al., 2014. Distributed File System For Load Rebalancing In cloud Computing. , 2, pp.15–20.

[20] Popović, K. & Hocenski, Z., 2015. cloud computing security issues and challenges. , (January), pp.344–349.

[21] Venkata Narasimha Inukollu , Sailaja Arsi and Srinivasa Rao Ravuri "Security issues associated with big data in cloud computing "International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014

[22] Tene, O. & Polonetsky, J., 2012. Privacy in the Age of big data.