Review Paper on Big Data Analytics in Cloud Computing

Saneh Lata Yadav[1], Asha Sohal[2]
Assistant Professor, Department of Computer science and Engineering
K. R. Mangalam University, India
Associate Professor, Department of Computer science and Engineering
K. R. Mangalam University, India

Abstract: *Cloud computing is a most powerful technology which performs massive-scale and complex computing. It eliminates the requirement to maintain costly computing hardware, dedicated space requirement and related software. Massive growth in the scale of data or big data generated through cloud computing has been identified. Concept of big data is a challenging and time-demanding task that requires a large computational space to ensure successful data processing and analysis. This paper includes definition, characteristics, and classification of big data along with some discussions on cloud computing are introduced. The similarities between big data and cloud computing, big data storage systems, several big data processing techniques and Hadoop technology are also discussed. The term 'Big Data' defines innovative techniques and technologies to capture, store, distribute, manage and analyze petabyte-or larger-sized datasets with high-velocity and different structures. Big data may be structured, unstructured or semi-structured, resulting in incapability of conventional data management methods. Data can be generated from various relevant sources and can store in the system at various rates. In order to analyze these large amounts of data in an inexpensive and efficient way, parallelism technique is used. 2015 was the year that Big Data went from being something that a majority of organizations were either doing or at the very least actively considering. The growth of cloud-based Big Data services has made Big Data analytics a feasible reality for organizations of all sizes.*

Keywords: *Big Data, Big Data Analytics, Map Reduce, Hadoop.*

I. INTRODUCTION

Big data is a word used for detailed information of massive amounts of data which are either structured, semi structured or unstructured. The data which is not able to be handled by the traditional databases and software Technologies then we divide such data as big data. The term big data is originated from the web companies who used to handle loosely structured (numerical form, figures, and transaction data etc.) or unstructured data (Email attachments, Images comments on social networking sites) [1]. The big data is defined using five V's. Volume includes many factors contribute for the increase in volume like storage of data, live streaming etc. Variety consists of various types of data is to be supported. Velocity means speed at which the files are created and processes are carried out refers to the velocity. Veracity indicates data reliability with respect to big data exploitation. Value shows worth with respect to big data exploitation. Since big data is not only large but also different and fast-growing. Some analytical techniques are required in order to the attempt some relevant information. It gives a broad overview of some of the most commonly used techniques and technologies to help the reader to better understand the tools based on big data analytics. There are many analytic techniques that could be employed when considering a big data project. Which ones are used that depends on the type of data being analyzed, the technology available to you, and the research questions you are trying to solve? Some of the tools that came up frequently in the reviewed material are summarized here. It is often used in data mining and according to Chen, Chiang, and Storey (2012) it lends support to recommender systems like those employed by Netflix and Amazon. Data Mining: Manyikaet al. (2011) calls data mining "combining methods from statistics and machine learning with database management" in order to pinpoint patterns in large datasets [2]. Picciano (2012) lists it as one of the most important terms related to data-driven decision making and describes it as "searching or 'digging into' a data file for information to understand better a particular phenomenon." Crowd sourcing collects data from a large group of people through an open call, usually via a Web2.0 tool. This tool is used more for collecting data than for analyzing it. Machine learning includes traditionally computers only know what we tell them, but in machine learning, a subspecialty of computer science, we try to craft "algorithms that allow computers to evolve based on empirical data. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data" (Manyika et al.

2011). Miller (2011/2012) gives the example of the U.S. Department of Homeland Security, which uses machine learning to identify patterns in cell phone and email traffic, as well as credit card purchases and other sources surrounding security threats. They use these patterns to try to identify future threats so they can handle them before they become large problems. A large portion of generated data is in text form. Emails, internet searches, web page content, corporate documents, etc. are all largely text based and can be good sources of information. Text analysis can be used to extract information from large amounts of textual data. This can be done to model topics, mine opinions, answer questions, and other goals.

II. RELATED WORK

With the help of analytical techniques, there are several software products and many technologies to facilitate big data analytics. Some of the most common will be described in this paper. Enterprise data warehouses are databases used in data analysis [3]. Russom (2011) writes that for many popular businesses that are taking step to start handling big data the big question is that can the present or planned enterprise based data warehouse (EDW) handle big data and advanced data analytics without degrading performance of other workloads for reporting and online analytic processing? Some popular institutions manage their analytic data in the EDW by its own while others use a different platform, which helps relieve some of the burden on the server resulting from managing your data on the EDW [4]. Many new visualization products aim to fill this need, dividing methods for representing data points numbering up into the millions. Russom (2011) shows this field as one of those having the most potential and says it is poised for aggressive adoption. Beyond simple representation visualization can also involve in finding the information search. Hansen, Johnson, Pascucci, and Silva wrote an article included in Hey, Tansley, and Tolle's collection. The fourth paradigm (2009) telling visualization in data-intensive science in which they define that visualization products allow us to compare models and datasets. It enables quantitative and qualitative decision-making and their article focuses scalability in visualization technologies and their ability to track provenance in real-time [5].

A. *HADOOP*

This is a most available java based programming framework which supports the processing of large amount of data in a distributed computing environment. With the help of Hadoop, big amount of data sets can be analyzed over cluster of servers and applications can be run on system with thousands of nodes involving terabytes of information as shown in fig 1.

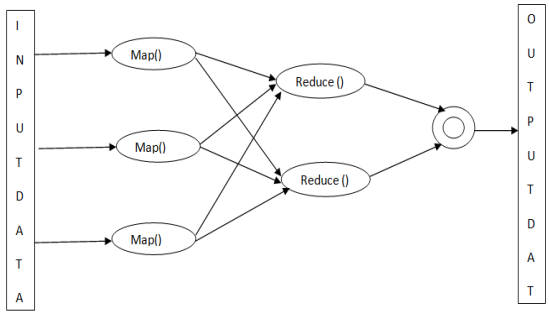

Fig. 1: Hadoop Structure

This decreases the risk of system failure even when a large amount of nodes fails. It includes a scalable, flexible, fault tolerant computing solution. HDFS defines a file system spanning all nodes in a Hadoop cluster for data storage connects the file systems on local nodes to make it onto a very large file system thus improving the reliability [6]. Task trackers are responsible for executing the tasks that the job tracker assigns them. Job trackers have two major responsibilities which are managing and controlling the cluster resources and then schedule all user jobs. Data engine consists of all the information about the processing the data. Fetch manager protects and fetch the data while particular task is running.

B. *MAP REDUCE*

Map Reduce [7] framework is basically used to write apps that analyze large amounts of data in a manner of reliable and fault tolerant. Initially the application is divided into individual chunks which are analyzed by individual map jobs by following the concept of parallelism. The result of map sorted by a framework and then sent to the reduce tasks. The supervision is taken care by the framework. The framework splits the data into smaller chunks that are processed in parallel on cluster of machines by programs called mappers. The result from the mappers is then consolidated by reducers into desired result as shown in fig 2.

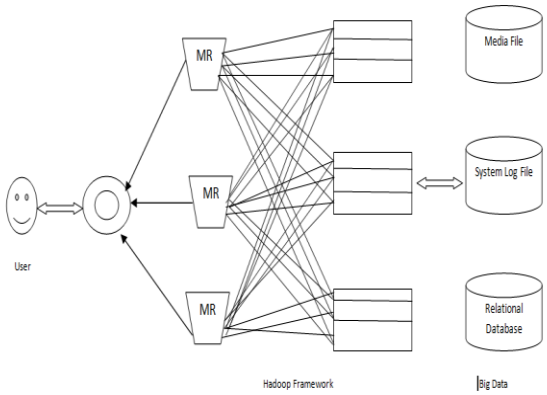

Fig. 2: Map Reduce

The share nothing architecture of mappers and reducers make them highly parallel [9]. Over the last many years, there are so many researchers has completed their work successfully on big data. Hundreds of articles have appeared in the general business press (For example Forbes, Fortune, Bloomberg, Business week, The Wall street journal, The Economist)[15]. National Institute of Standards and Technology [NIST] said that Big Data in which data volume, velocity and data representation ability to perform effective analysis using traditional relational approaches [16]. In March 2012, The Obama Administration great researcher announced that the US would spend 200 Million Dollars to launch a big data research plan [17]. An IDC diaries predicts that from 2005 to 2020, the global data volume will increase by a factor of 300, from 130 Exabyte's to 40,000 Exabyte's, showing a double growth every two years[18]. IBM gives estimation that everyday 2.5 quintillion bytes of data are generated out of which 90% of the data in the world today has generated in the last two years. It is analyzed that social networking sites like Facebook have 850 Million users, LinkedIn has 110 million users and Twitter has 350 million users [19]. From industry, government and research community, it is predicted that Big Data has led to an emerging and recent research field that has attracted tremendous interest of users. The major interest is first exampled by coverage on both industrial reports and public media [20]. For example, Mobile Phones becoming best way to get data from people in different aspect, the large amount of data that mobile carrier can process to improve our daily life [21]. In figure 3, From Year 2005, it would show from this graph that the large amount of data was practically increased. However, Consider exponential growth in data from 2005 year, when considering enterprise system and user level data was flooding into data warehouse [22]. Data was in structured form when it creates from many organizations. Data goes from three properties like volume, Variety and velocity. Many companies were suffering from the problems on how to expand the capacity of data warehouse to accept and create new requirement.

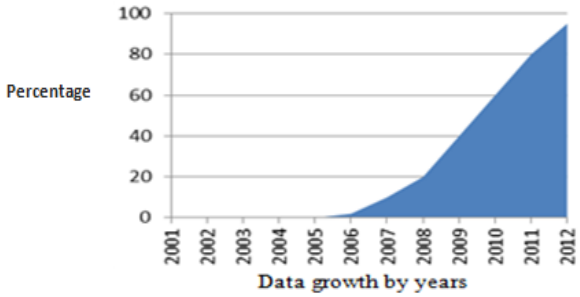

Fig 3: Data Growth

S. Vikram Phaneendra & E. Madhusudhan Reddy Illustrated that in olden days the data was less and easily handled by RDBMS but recently it is difficult to handle huge data through RDBMS tools, which is preferred as "big data". In this they told that big data differs from other data in 5 dimensions such as volume, velocity, variety, value and complexity. They illustrated the Hadoop architecture consisting of name node, data node, edge node, HDFS to handle big data systems. Hadoop architecture handle large data sets, scalable algorithm does log management application of big data can be found out in financial, retail industry, health-care, mobility, insurance. The authors also focused on the challenges that need to be faced by enterprises when handling big data: - data privacy, search analysis, etc [10].

Kiran kumara Reddi & Dnvsl Indira enhanced us with the knowledge that Big Data is combination of structured, semi-structured ,unstructured homogenous and heterogeneous data .The author suggested to use nice model to handle transfer of huge amount of data over the network .Under this model, these transfers are relegated to low demand periods where there is ample ,idle bandwidth available . This bandwidth can then be repurposed for big data transmission without impacting other users in system. The Nice model uses a store and forward approach by utilizing staging servers. The model is able to accommodate differences in time zones and variations in bandwidth. They suggested that new algorithms are required to transfer big data and to solve issues like security, compression, routing algorithms [11].

Wei Fan & Albert Bifet Introduced Big Data Mining as the capability of extracting Useful information from these large datasets or streams of data that due to its Volume, variability and velocity it was not possible before to do it. The author also started that there are certain controversy about Big Data. There certain tools for processes. There are certain Challenges that need to death with as such compression, visualization etc. [12].

Albert Bifet Stated that streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge, allowing organizations to react quickly when problem appear or detect to improve performance. Huge amount of data is created everyday termed as "big data". The tools used for mining big data are apache hadoop, apache big, cascading, scribe, storm, apache hbase, apache mahout, MOA, R, etc. Thus, he instructed that our ability to handle many Exabyte's of data mainly dependent on existence of rich variety dataset, technique, software framework [23].

Bernice Purcell started that Big Data is comprised of large data sets that can't be handle by traditional systems. Big data includes structured data, semi-structured and unstructured data. The data storage technique used for big data includes multiple clustered network attached storage (NAS) and object based storage. The Hadoop architecture is used to process unstructured and semi-structured using map reduce to locate all relevant data then select only the data directly answering the query. The advent of Big Data has posed opportunities as well challenges to business [24].

III. CHALLENGES AND OPPORTUNUITIES IN BIG DATA

We live in the period of the big data where we can collect more and more information from daily life of human being. So far, researchers are failed to unify the features that are more essential to big data, many think that big data is something which we cannot process or analyze using existing technology, theory or any other method of such kind. However the world has become helpless since enormous amount of data is being generated by science, business, social sites and even society. Big data has posed many challenges to the IT industry [8].

IV. CONCLUSION

This paper gave a description of a systematic flow of survey of the big data in the environment of cloud computing. Big data is the large and complex datasets and it is created from various sources like social media likes, comments, playing a video game, email attachments etc. There is complexity in big data such as velocity, variety and volume. These three terms are more challenging for big data. We have also seen some technologies and techniques. Since big data is not only large, but also varied and fast-growing many technologies and analytical techniques are needed in order to attempt extracting relevant information. The benefits are many and varied, ranging from higher quality education to cutting-edge medical research, and while further research is required for things like ensuring people's information is protected from exploitation, there are many exciting and innovative discoveries waiting to be uncovered through big data analytics. It is very much required that the computer scholars and IT professionals to cooperate and make a successful and long term use of cloud computing and explores new ideas for the usage of the big data over cloud environment.

REFERENCES

[1] A. Abouzeid, K. B. Pawlikowski, D. J. Abadi, A. Rasin, and A. Silberschatz. HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. PVLDB,2(1):922–933, 2015.

[2] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony,H. Liu, P. Wyckoff, and R. Murthy. Hive - A Warehousing Solution Over a Map-Reduce Framework. PVLDB, 2(2):1626–1629, 2009.

[3] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.

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

[5] Wie, Jiang , Ravi V.T, and Agrawal G. "A Map-Reduce System with an Alternate API for Multi-core Environments.". Melbourne, VIC: 2010, pp. 84-93, 17-20 May. 2010.

[6] F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Busines on big dataapplications." Big Data, 2013 IEEE International Conference, Silicon Valley, CA, Oct 6-9, 2013, pp.32 – 37.

[7] Xu-bin, LI , JIANG Wen-rui, JIANG Yi, ZOU Quan "Hadoop Applications in Bioinformatics." Open Cirrus Summit (OCS), 2012 Seventh, Beijing, Jun 19-20, 2012, pp. 48 – 52.

[8] 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.

[9] Elragal, A. (2014). ERP and Big Data: The Inept Couple. Procedia Technology, 16, 242-249.

[10] S.Vikram Phaneendra & E.Madhusudhan Reddy "Big Data-solutions for RDBMS problems- A survey" In 12th IEEE/IFIP

Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

[11] Kiran kumara Reddi & Dnvsl Indira "Different Technique to Transfer Big Data : survey" IEEE Transactions on 52(8) (Aug.2013) 2348 { 2355}

[12]Umasri.M.L,Shyamalagowri.D,SureshKumar.S"Mining Big Data:- Current status and forecast to the future" Volume 4, Issue 1, January 2014 ISSN: 2277 128X.

[13] Albert Bifet,"Mining Big Data in Real Time", informatica, 2013.

[14] James Manyika, Michael Chui, Brad Brown, Jacques Bhuhin, Richard Dobbs, Charles Roxburgh, Angela Hungh Byers, "Big Data: The next frontier for innovation, competition and productivity", June 2011.

[15] Sameera Siddiqui, Deepa Gupta," Big Data Process and Analytics : A Survey", International Journal Of Emerging Research in Management & Technology, ISSN: 2278-9359, Volume 3, Issue 7, July 2014.

[16] M.Cooper, P.Mell(2012). Tackling big Data(Online).Http://csrc.nist.gov/groups/SMA/Forum/document/June 2012Presentation/f%CSM_june2012_cooper_Neul.pdf.

[17] Han Hu, YongyangNen, Tat Seng Chua, Xuelong Li," Towards Scalable System for Big Data Analytics: A Technology Tutorial", IEEE Access, Volume 2, Page No 653, June 2014.

[18] J.Gantz, D. Reinset," The Digital Universe in 2020: Big Data, Bigger digital shadow, and biggest growth in the far east", in Proc : IDC iview, IDC Anal, Future, 2012.

[19] www.ebizmba .com/articles/social-networking-websites.

[20] Neil Raden,"Big Data Analytics Architecture", Hired Brains Inc, 2012.

[21] James Manyika, Michael Chui, Brad Brown, Jacques Bhuhin, Richard Dobbs, Charles Roxburgh, Angela Hungh Byers, "Big Data: The next frontier for innovation, competition and productivity", June 2011.

[22]Wei Fan, Albert Bifet, "Mining Big Data: Current Status and Forecast to the Future", SIGKDD Explorations, Volume 14, Issue 2.

[23] Albert Bifet"Mining Big Data In Real Time" Informatics 37 (2013) 15–20 DEC 2012.

[24] Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013.

[25] Ritu Katara, Hareram Shah "A Novel Integrated Approach for Big Data Mining", International Journal For Computer treands and Technology, Volume 18, Number 5, Dec 2014.

[26] M. Saranya, A. Prema "Survey on Big Data Analytics Using Hadoop ETL", International Journal For Computer treands and Technology, Volume 48, Number 5, June 2017.

[27] V. Harsha Shastri,, V. sreeprada, T. Kavitha "A Survey on Big Data Technologies, Challenges and Impact on Internet of things", International Journal For Computer treands and Technology, Volume 35, Number 3, May 2016.

[28] Fayaz Ahmad Lone, dr. Amit Kumar Chaturvedi "Proposing a Novel Model on Security Challenges in Cloud Computing especially Social Media and social Sites", International Journal For Computer treands and Technology, Volume 47, Number 1, May 2017.