International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

Systematic Review: Progression study on BIG DATA articles

Vaishali Chauhan[1] Dr. Munish Sabharwal[2]

[1]Student of Doctor of Philosophy, Department of Computer Science and Engineering, Chandigarh University, (India)
[2]Head of department Department of Computer Science and Engineering, Department of Computer Science and Engineering, Chandigarh University, (India)

ABSTRACT - Data extraction has gotten significant consideration because of the fast development of organized, unstructured and semi-organized information. The specialist needs a minimal effort, adaptable, simple to-utilize and adaptation to non-critical failure stage for huge volume information handling excitedly. This investigation has the target of methodical audit towards exhibiting a dream on information examination done on enormous information.

Keywords: Big Data, Hadoop, Performance evaluation

I. INTRODUCTION

Big data depicts the propinquity between data size and data preparing speed in a framework. Data has been a spine of any undertaking and will do as such pushing ahead. Putting away, extricating and using data has been vital to many organization's tasks. In the past when there were no interconnected frameworks, data would remain and be expended in one spot. With the beginning of Web innovation, capacity and prerequisite to share and change data have been a need. The blast of data is being experienced by each segment of the figuring business today. Web mammoths, for example, Google, Amazon, Facebook and such need to manage tremendous measures of client produced data as blog entries, photos, status messages, and sound/video documents. Previously, the kinds of data accessible were constrained. Way to deal with innovation has a well-characterized set of Data the executives. Be that as it may, in this day and age, our reality has been the blast of data volumes. It is Terabytes and petabytes. The enormous existing data from various databases social data put away in Big Data. The majority of this data would be pointless on the off chance that we couldn't store it, and that is the place Moore's Law [2] comes in. The law which expresses that the quantity of transistors on incorporated circuits pairs at regular intervals, since the mid '80s processor speed has expanded from 10 MHz to 3.6 GHz - an expansion of 360 (not including increments in ―word length‖ and number of centers)? In any case, we've seen a lot bigger increments away limit, on each dimension. Smash as moved from $1,000/MB to generally $25/GB—a cost decrease of around 40,000, and this together with the decrease in size and increment in speed. The principal gigabyte plate drives showed up in 1982, gauging in excess of 100 kilograms; presently terabyte drives are purchaser gear, and a 32 GB small scale SD card weighs about a large portion of a gram. Regardless of whether you take a gander at bits for every gram, bits per dollar, or crude limit, stockpiling accessibility has become quicker than CPU speed.

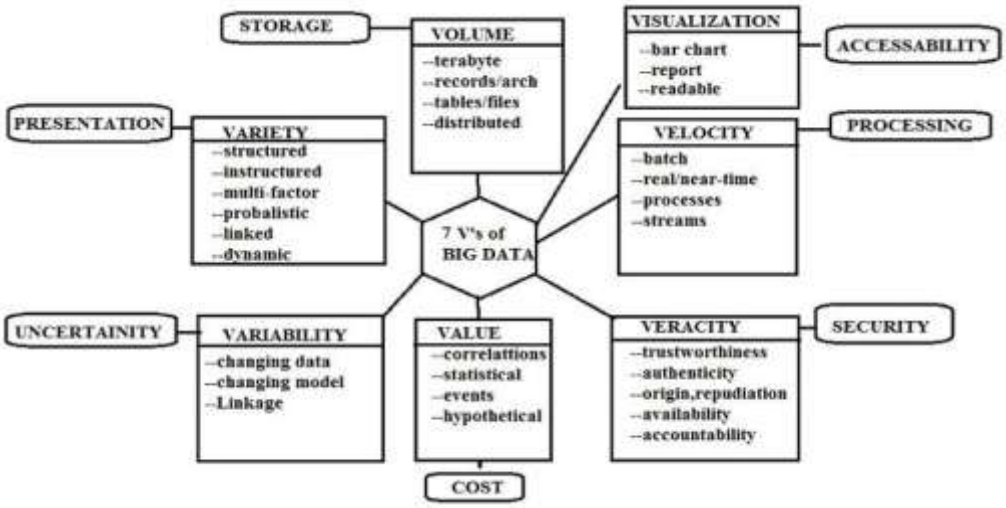
In the accompanying table, we give the units used to quantify the data, beginning with those most natural to those that are basic to gauge the Big Data.

Table 1: Units of data

Name	Symbol	Value in bytes
Gigabyte	GB	10^9(=1000^3)
Terabyte	TB	10^{12}(=1000^4)
Petabyte	PB	10^{15}(=1000^5)
Exabyte	EB	10^{18}(=1000^6)
Zettabyte	ZB	10^{21}(=1000^7)
Yottabyte	YB	10^{24}(=1000^8)

The term itself is in effect more formally characterized by IBM as the mix of 3 V's is speed, assortment, and volume. These are the conventional big data properties. Notwithstanding, the gained properties delineated in the wake of entering the framework incorporates esteem, veracity, inconstancy, and perception. Subsequently, 7 V's effectively portrays the big data [3][4].

- **Volume:** A great many data is transferred each day on Facebook, Twitter, and other online stages. The framework is producing terabytes, Petabyte and zeta bytes of data. This tremendous piece of data is dealt with through the procedure of Data Mining growing its extension to cover big data examination.



- **Velocity:** The framework creates floods of data and numerous sources that necessitate that data. There is an exponential development in data consistently. Consistently the data is overflowed with a large number of online transfers. The generally acknowledged databases have expanded to millions requiring highlights choice as a crucial necessity. Different CI methods are utilized for time-space cosmology (TDA) [5].

- **Variety:** Big data is a piece of structures and unstructured data which incorporate online journals, pictures, sound, and recordings. These data might be dissected for slant and substance. Prior might be the days when organizations managed just a solitary data group however today big data gives a stage to all data designs.

- **Variability:** Big data permits taking care of vulnerability in data with changing data helping in the expectation of future conduct of different clients, business visionaries, and so forth. Fundamentally, the significance of data is always showing signs of change and the data depends for the most part on language handling.

- **Veracity:** So as to guarantee the exactness of big data different security instruments are accommodated guaranteeing potential estimation of the data. This includes mechanized basic leadership or nourishing data into an unsupervised AI calculation. This guarantees the realness, accessibility, and responsibility of the data.

- **Visualization:** The systems engaged with making the data coherent and effectively available commitment to the fifth V of the Big data. The data should be effectively comprehended and the systems, for example, the different enhancement calculations give a preferred position of giving an ideal audit of the data dissected.

- **Value:** The estimation of big data is colossal. It empowers conclusion investigation, forecast, and proposal. It is monstrous and quickly extending, yet it loses its value when managed without examination and representation that experiences uproarious, chaotic and quickly evolving data.

2. HADOOP

Hadoop is intended to scale up from single servers to a great many machines, each offering nearby calculation and capacity. As opposed to depend on equipment to convey high-accessibility, the library itself is intended to distinguish and deal with disappointments at the application layer, so conveying an exceedingly accessible administration over a group of PCs, every one of which might be inclined to failures|| [6]. Hadoop was at first motivated by papers distributed by Google, delineating its way to

deal with taking care of a torrential slide of data, and has since turned into the standard for putting away, handling and examining several terabytes, and even Petabyte of data. Hadoop structure improvement was begun by Doug Cutting and the system got its name from his child's elephant toy [7]. Hadoop should have the boundless scale-up capacity and hypothetically, no data is too big to deal with conveyed design [8].

The two most important components that are the foundation to Hadoop framework are:

❖ **Hadoop Distributed File System – HDFS**

HDFS is a dispersed document framework intended to keep running on item equipment. HDFS has an ace/slave design. See Figure 3.1. It's a compose once and perused on various occasions approach. Hadoop Circulated Document Framework (HDFS) is a record framework which is utilized for putting away enormous datasets in a default square of size 64 MB in the conveyed way on Hadoop group [9]. A HDFS bunch comprises of a solitary NameNode (most recent variant 2.3.0 has repetitive NameNode to maintain a strategic distance from the single purpose of disappointment), an ace server machine that deals with the document framework and manages access to the filesystem by the customers. There are different data hubs per bunch. The data is part of squares and put away on these data hubs. NameNode keeps up the guide of data dissemination. Data Hubs are in charge of data perused and compose tasks amid the execution of data examination. Hadoop additionally pursues the idea of Rack Mindfulness. This means a Hadoop Director client can characterize which data lumps to save money on which racks. This is to anticipate the loss of the considerable number of data if a whole rack falls flat and furthermore for better system execution by abstaining from moving big lumps of cumbersome data over the racks. This can be accomplished by spreading duplicated data obstructs on the machines on various racks. Figure 3 [10], the NameNode and DataNode are product servers, normally Linux machines. Hadoop runs distinctive programming on these machines to make it a NameNode or a DataNode. HDFS is fabricated utilizing the Java language. Any machine that Java can be kept running on can be changed over to go about as the NameNode or the DataNode. A run of the mill bunch has a devoted machine

that runs only the NameNode software. Each of the other machines in the cluster runs one instance of the DataNode software. The NameNode manages all HDFS metadata [11].

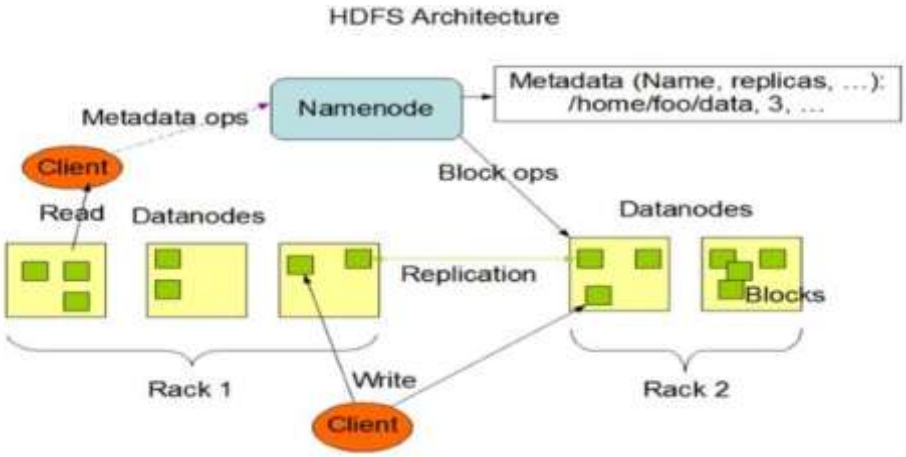


Fig2: Hadoop Architecture [10].

❖ **Map Reduce Architecture**

It is a programming framework for distributed computing, which was created by Google using divide and conquer method to crack complicated Big Data problems into small units of work and process them in parallel [12]. The basic meaning of Map Reduce is dividing the large task into smaller chunks and then deal with them accordingly, thus, this speed up the computation and increase the performance of the system. Map Reduce can be divided into two steps:

❖ **Map Stage**

Map is a function that splits up the input text, so map function is written in such a way that multiple map jobs can be executed at once, map is the part of the program that divide up the tasks [13]. This function takes key/value pairs as input and generates an intermediate set of key/value pairs.

❖ **Reduce Stage**

Reduce is a function that receives the mapped work and produces the final result [14]. The working of Reduce function depends upon merging of all intermediate values associated with the same intermediate key for producing the final result.

3. METHODOLOGY

The efficient survey was critical to decide from the earliest starting point of the convention to be pursued. The reason for the exploration was IEEE Explorer and Google Scholar, search keyword "Analysis in Big Data".

After the underlying exploration, 64 articles were found on the two consulted databases of movement

TABLE 2: Progression study on BIG DATA articles

Progression	IEEExplore	Google Scholar
Year exposed	2014 -2018	2014 - 2018
Articles downloaded	32	32
Real Database	8	9
historical data	10	12
Search keywords	"Analysis in Big Data"	"Analysis in Big Data"
Experimentation Articles	9	4
Case study based articles	5	7
Finally selected articles	8	11

4. RESULTS AND DISCUSSION

For this research, 64 articles were analyzed, all of them published after 2014

Table 2 represents a progressive number of published articles on Big Data over the years. The first article to be analyzed is from 2014. In the following years, the number of publications increased, as we can see in 2018. The experimental and case study article where 25 articles out of the experimental articles in google scholar where 4 articles and IEEExplore where 9 articles. In case study articles in google scholar where 7 articles and IEEExplore where 5 articles. The final selected article 19 , out of which Google Scholar 11 and IEEEXplore 9.

5. CONCLUSION

The intention of this study was to bring an expansive image of the cutting edge about Big Data. As per the examination led, it was conceivable to anticipate that, in spite of being another subject, the quantity of distributions about the topic is expanding, what demonstrates the significance and the enthusiasm on the issue. For future work, it is crucial to think about Hive and Pig on bigger data accumulations and giving out the best outcomes out of the two.

REFERENCES

1. Jacobs, The pathologies of big data, Commun. ACM Vol. 52 (8) (2009) pp. 36–44.
2. R. Schaller, Moore's law: past, present and future in Spectrum, IEEE Vol. 34 (1997): 52-59 (availablehttp://mprc.pku.edu.cn/courses/organization/autumn2013/paper/Moore%27s%20Law/M132oore
3. %27s%20law%20past,%20present%20and%20future.pdfaccessed December 2013).
4. Rasmus Wegener and Velu Sinha, -The Value of Big data: How analytics differentiates winners||, Bain and Company.
5. Yaochu Jin,, Barbara Hammer, – Computational Intelligence in Big Data||, IEEE Computational intelligence magazine, August 2014.
6. Huijse et al.,|| Computational intelligence challenges and Applications on Largee Scale Astronomical Time Series

Databases||, IEEE, 2013.

7. Apache Hadoop. What Is Apache Hadoop?, 2014. http://hadoop.apache.org/, accessed April 2014.

8. Wikipedia. Apache Hadoop, 2014. http://en.wikipedia.org/wiki/Apache_Hadoop, accessed April2014.

9. T. White. Hadoop – the definitive guide. O'Reilly Media, Inc., Sebastopol, California, 1 edition, 2009.

10. Qureshi, S. R., & Gupta, A, —Towards efficient Big Data and data analytics: A review||, IEEE International Conference on IT in Business, Industry and Government (CSIBIG),March 2014 pp-1-6.

11. Apache Hadoop. MapReduce Tutorial, 2013. https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html, accessed April 2014.

12. Apache Hadoop. HDFS Architecture Guide, 2013. http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html, accessed April 2014.

13. Aravinth, M. S., Shanmugapriyaa, M. S., Sowmya, M. S., & Arun, —An Efficient HADOOP Frameworks SQOOP and Ambari for Big Data Processing,|| International Journal for Innovative Research in Science and Technology, 2015, pp. 252-255.

14. Liu, Z., Yang, P., & Zhang, L. (2013). A sketch of big data technologies. In 2013 seventh international conference on internet computing for engineering and science (pp. 26–29).

15. Ramesh, B. (2015). Big data: Architecture (vol. 11). India: Springer India.

16. Ward, J. S., & Barker, A. (2013). Undefined by data: A survey of big data definitions. arXiv.org, 1, 2.

17. Benjelloun, F.-Z., Lahcen, A. A., & Belfkih, S. (2015). An over- view of big data opportunities, applications and tools. In Intelligent systems computer vision (ISCV), 2015 (pp. 1–6).

18. Maria, R. E., et al. (2015). Applying scrum in an interdisciplinary project using big data, internet of things, and credit cards. In Proceedings of 12th international conference on information tech- nology: New generations ITNG 2015 (pp. 67–72). Linz.

19. Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. January, 19, 171–209.

20. Schneider, R. D. (2012). Hadoop for dummies (1st ed.). EUA: John Wiley & Sons.