www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 11 November 2017, Page No. 23083-23088
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i11.14

Big Data Analysis: A Review

Chaitanya Singh* & Md. Arsh Saifi*
(*Students)
Department of Computer Science and Engineering
Dronacharya Group of Institutions
Greater Noida, UP

Abstract
Name defines its size only besides, broader dimensions and scopes define Big Data in most appropriate terms. This paper attempts to provide a consolidate knowledge of Big Data by covering its peripheral definitions and benefits. With the advancement in technologies and enhanced scale of Internet of Things, millions of people get the avail services and customize them accordingly. Besides, there emerge needs for analyzing the information gained, arranging the data and retrieving useful information. Here Big Data and its opportunities can be visualized.

Data analysis is the root and essential step for retrieving Big Data and their proper processing. This paper provides fundamental and ground level knowledge about Big Data, its beneficial phases, challenges, and its essential role for creating vast positive differences in many fields from social networking to e-commerce to medical sciences and in many other scenarios.

Many researches are made under this title showing its worth and scope besides, there are many other researchers indulged in developing methods and processes to more and more effective computation of Big Data. This review thesis gives away a comprised and multi-defined prospect of Big Data covering all its related knowledge and concepts. Traditional aspects of analysis of Big Data is also covered along with the latest technology discussion of Hadoop and further its limitations. This paper also concludes and reinforces the desired necessity for to innovate new tools and mechanisms for relevant computation and analysis of structured as well as unstructured Big Data. Therefore, for deceiving the complexity and incrementing the opportunities of Big Data in order to visualize enormous useful integrated phase for advancement, this paper attempts to map primarily bottom up approach to learn and understand Big Data in more precise manner.

Keywords: Big Data, Traditional Analysis techniques, Vs of Big Data, Hadoop, Privileges by Big Data, Challenges by Big Data, Dimensions of Big Data.

Introduction: Big Data is a crucial parametric upcoming center of focus for modern science, statistics, business and organization of management. [1]The concept of "Big Data" is nascent and moreover, its evolution and origin is uncertain. Big Data comprises of very single parameters starting from click stream data to every electronic and protonic information of advanced sciences and genomic stats of biological sciences and medicines.

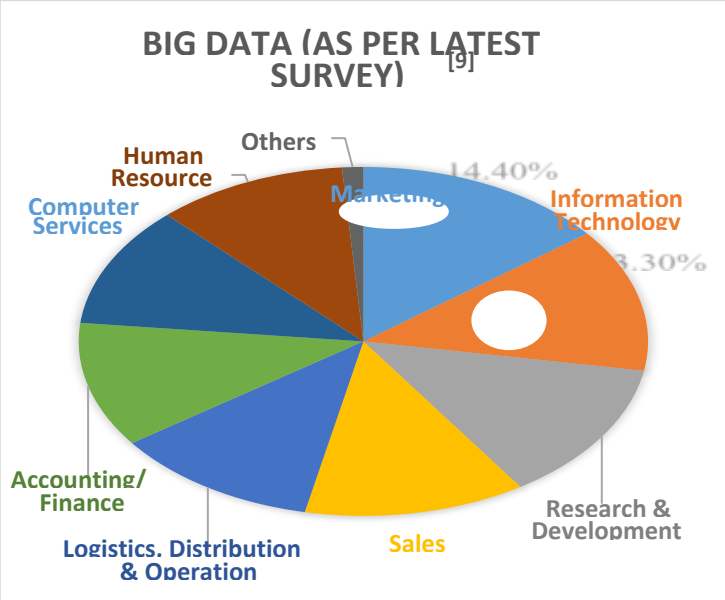
For an example, a trending video is shared by many Facebook users, then one shared video is then shared by some other people and then this cycle gets continued for endless time. So, this is the practical view of what is the dimension of Big Data and how there is a huge need for its regulations, e.g. as in the previous case, the record of users who shared the video is complicate enough to compute but on the other side worthily needed. Similarly window shopping demands for proper channelizing of every visit on their shopping sites, every look and catches of their customers, each filters, demands, trends, etc. and for these parameters to be easily computed proper segregation of Big Data is needed. Besides the need, handling of Big Data is another limiting scope as its concept is enormously wide and complex.

Big Data is playing the base for innovations and discoveries as well as analyses for economic value growth. The year 2017 is going to be distinguished itself in many parameters. Few glances can be : 95% of business professionals supporting personalization of customer experience, $30 million annual savings by controlling social media data in claims and frauds analytics, organizations to be relied on smart data more than on big data, many businesses are investing more than 300% in big data analysis, 28% rate of

growth of relevant source of unstructured data, around 43% of customer service teams that don't have real- time analytics will continue to shrink, many such more.
This statistics are truly stating the real and wider view of role of Big Data in present as well as in upcoming days. Moreover, the coming efforts and researches need to be in approach of Smart Data. Smart Data is derived from Big Data only, supervising more quality information for easily implementable computing process.

Big Data: Big Data provides many paths for marking heterogeneities which can thereby helpful for getting further advanced and accurate results, as not possible in small scale data. Similarly on the cons side, enormous size and dimension of Big Data demands for distinguished handling approaches to deal with the causing challenges which are like: scale and storing difficulties, noise production, correlating links, errors of measurement and generic deviations.
[2]The concept of big data came in executive shore via IT and various other engineering departments. At the core, big data refers to massive quantity of information from various sources being fed into data stores on a timesensitive basis. Big data is an important foundation of quality business intelligence, providing enough information to detect meaningful trends.
Big data is generally stored in databases, and is analyzed using various programming software which are specially designed to handle large, complex data sets. Data analysis concern at the relations for different types of data, such as demographic data and history of purchases, in order to examine and determine whether correlation exists.
While most people managed to work with megabytes, gigabytes and terabytes, large companies or governmental agencies may have to manage petabytes (1,000 terabytes) or exabytes (1,000 petabytes) or even more with passing days.

Sources of Big Data [3]: Data in excessive and massive amount causing high terms of management and complexity in processing and retrieval. Besides, studying



and analyzing the Big Data, it is also important to know and gather information related to the sources from where there are derived in order to retrieve efficient and trustworthy information for more advanced approach of analysis.

Social networking: Demographic information get collected in surplus and massive amount from different social networking platforms like: LinkedIn, Facebook, Twitter, Google, Yahoo and many individual-specified sites of travelling, gaming, shopping, etc.

Data as generated by activity: Internet of things i.e. computer and mobile device log files. Many other sensor data which includes check-ins and location tracking related information.

E-commerce and business statistics: Many commercial and shopping sites always want to keep an eye on customers' demands and choices. For these they gather data and information varying from a click to view of desired product to customers' reviews, etc.

Goals of Big Data:

- Evolution of suggestive and prominent methods for analyzing Big Data, so for feasible observations and precise predictions.
- Along with establishing links between characteristics and results for desired scientific motives.
- Also to compute the heterogeneities in widely enormous sample.
- Besides computing the common responses in different sub-divided samples.

Privileges by Big Data:

- Big data is a time- saving approach: For more than 60% time spam of working hours in a day, workers spend their efforts and knowledge to find and manage big data.
- Big data is Accessible: Many employers, even senior executives, clam that accessing the desired data is next to difficult.
- Big Data is Holistic: Data is, nowadays, kept in silos in an organization. So, their characteristic features are intimately interconnected with respect to whole.
- Big Data is Trustworthy: About 30% of companies capital measures for poor data quality and hence, suffer with backwardness in updating of information. This in return big data can save millions of dollars.
- Big Data is relevant: Around 45% of companies are dissatisfied with their methods and tools for segregating and extracting irrelevant information from massive data batches. This can influence tons of insight into acquisition efforts.
- Big Data is secure: The average data security breach costs around $200 per customer. 1.5% of annual

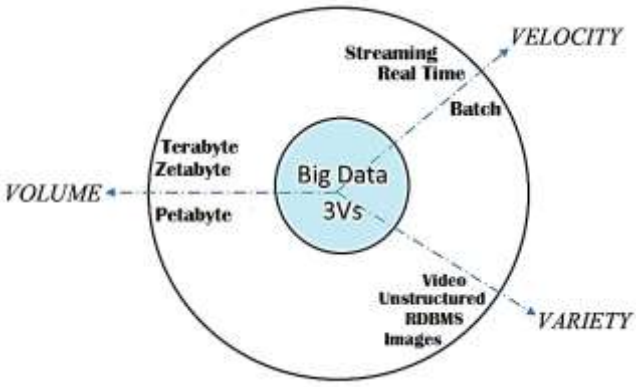
revenue of a company can be save by providing secure platforms built by hosts of big data.

- Big Data is Authoritive: 80% of companies and organizations suffers for finding multiple versions of facts and truths related to the source of their big data.
- Big data is Actionable: Old and outdated data results in more than 50% of companies making deconstructive forecast, predictions and bas decisions which can cost billions.

Five Dimensions of Big Data [4]**:** Big Data often when termed as "Smart Data" has a much broader scope in 5 dimensions: technology, application, economic, legal and social.

- Technology: There is demand for scalable platforms for data analysis methods to support overcome the skills gap. For example; enabling data analysis methods be accessible to a massive audience.
- Application: Many prestigious applications are emerging in the information business which extract and sell advanced and information enriched data. For examples; personalized medicine, industry 4.0 and digital humanities.
- Economic: The challenges and opportunities in the economic dimension lie in emerging business models and content delivery paradigm shifts. For example: information pricing and the role of open source software.
- Legal: From a legal perspective of vision, big data will display many challenges respectively to ownership, liability and insolvency, in addition to prevalent issues, such as privacy and security.
- Social: Lastly, data driven innovation will have a profound impact on society as a whole with respect to social interaction, news and democratic processes, among others.

Vs of Big Data [5]**:** The architecture of Big Data is comprising of several characteristic parameters namely 10 Vs of Big data among which Volume, Variety and Velocity play a pivotal role.



1. *Volume*: As there is an exponential growth in Big Data, including not just data in text format besides images, videos, graphical interchange format, audios ,etc. of large sizes, so it is very common to have Terabytes and Petabytes of storage in enterprises. This sometimes creates more complex results as during analysis same data is evaluated from multiple angles, creating more complexed big volumes which represents Big Data.

Photographs on Facebook can give an instance of Volume view of Big Data. Facebook has more than 200 billion images, no wonder as having much more users than population of India, but on analyzing the sizes it will be beyond the imagination. Coming to the world of Apps, a general app has counts of 10 million installs on Android, except iOS, web and Window users' counts. This gives rise to the emergence of massive volume of data. Many more such views arise from many different industries namely from automobile to medical industries to IT industries, etc. So these are producing data in huge volumes, i.e. composing the volume factor of Big Data.

Moreover, the data phases are mainly given by social media and e-commerce as they have given enormously massive data than ever expected. Resulting in failures for IT managers and data handlers for censored data and data security.

2. *Variety*: This is the most complex sight for Big Data. As there are many different formats of data like excel, excel, mat, doc, text files and many more, even it can be mp4, mp3, exe, software update files, sensor files etc. So this gives the situational insight that handling these varieties of data is a major challenge for managers.

For many of the times, organizations and managerial bodies face challenges in arranging and gaining useful information from these distinguished varieties. These diversity of huge unstructured data comprises the form of Big Data.

3. *Velocity*: Velocity shows the rate of data crossing a particular terminal with respect to time.

Coming to the previous instance of Facebook, the users upload more than 800 million photographs, videos and write-ups per day on an average. So the last evaluated figure of 200 billion can be summed up in few weeks.

In general terms velocity provides the insight of how fast certain data is coming in or going out. This in coming or outgoing data creates multiple linkage as data arrives from a source, then goes for processing, storing, sharing and retrieving if needed in future.

Moreover, for safely sharing of data, more and more data are encrypted to be sent through firewall, and so many different analyzing techniques are associated to evaluate the patterns, composition and matter of data shared.

Besides, other challenging factor arises when the velocity differentiates its particular type. For more general concept velocity is of two types: Batch data flow velocity & Data streaming velocity. This again generates great demand of proper analyzing techniques and applications for dealing with both the types of data whether it comes in batches or regular flow of data occurs.

4. *Validity*: Validity signifies the preciseness of the Big Data to be analyzed. Validating Big Data for better implementation in different aspects. Good governance is a key to manage the fundamental consistency in massive volumes of Big Data.

5. *Visualization*: Many challenges are there for operating visualizing tools of Big Data in order maintain and gain the quality, consistency and effective response time. These are dependent factors on velocity, variability and variety of analyzed Big Data.

6. *Variability*: Variables are the results of occurrence of inconsistencies in data. While analyzing variables are importantly studied in order to optimize the Big Data more efficiently. The reasons for these variabilities are the occurrence of diverse dimensions in Big Data.

7. *Volatility*: Volatility deals with basic hospitalities of Big Data considering their age relevancies, the storage characteristic, behavior, etc. Volatility sometimes emerges as an influential factor due to effect of velocity and volume of Big Data.

8. *Veracity*: Veracity is a limitation to Big Data. Veracity refers to the confidence in data relevancy. With the increment in above mentioned factors, veracity decreases abruptly. Veracity depends on relevancy of source, its context and its quality of being meaningful.

9. *Vulnerability*: Security is another important concern during analyzing Big Data. As the massive volume of Big Data employs to large sizes of data, many security breaches occur due to which failure in consistency and originality are often seen.

10. *Value*: Value being the most important factor for handling Big Data. Value determines the purpose, approach, handling techniques and results required from the Big Data. Proper function is to be identified to get constructive results before analyzing.

Challenges by Big Data [6]: We always define big data as a volume but the reality case is entirely different. This term especially refer to the technology which includes tools and processes required by an organization to handle large amounts of data and facilities of storage. In general terms, having more data on one's potential customer allows companies to better tailor and supervise their products and marketing efforts in order to create the highest level of satisfaction and repeat and advance business.

On the other side, big data can be trouble for companies, as big data create overload and intersection as noise. Most important key is the determination of relevant data from big data. Structured data, which consists of numeric values, can be easily sorted and stored. Unstructured data, such as emails, videos, and text documents, require more refined techniques to be applied before it becomes useful. The increment in the quantity of data available creates opportunities and challenges both, few of the challenges are listed below:

1. For achieving tremendous boost, a specific language is required by Big Data with invention of programming language for better analysis.

2. The influence of volume factor is to be treated in order to distinguish the signal of data and meaningful information. The main obstruction is to identity appropriate data and retrieving the data for advanced application. Again the crucial drawback is distinguishing between quality data and junk data.

3. Reinvention and re-architecture of Big Data tools and techniques for more advanced mechanism.

4. The variable of time duration is also an important factor causing difficulty for not only real time evaluation but also for long durational casualities related to Big Data.

5. Architecture of IT: As world is changing in every minute, resulting in difficulties of collaboration and maintainence of proper mechanism for analyzing Big Data and gaining efficient results.

6. Security emerges as a common threat for Big Data handling as even the arrived data is to be secured for better predictions and results. Moreover, the encrypted data needs protection as faulty programs may be in the same stock of massive Big Data volume.

Traditional Solutions for Big Data:

With the influence of WWW (World Wide Web) in the early 1990s, Big Data became a major centered issue for segregation, indexing and querying the stock of Data for more relevant information and prediction. Since the challenges were growing from the very early stages of internet advancements, so there emerged many traditional analysis methods for Big Data.

Traditional Data Analysis was meant for concentrating, extracting, refining and segregating meaningful data from the stock of huge chaotic data. This is done by use of advanced and suitable statistical ways. This results to analyze huge first and second hand data. Further extent, resulting in identification of subject of matter to utilize the bulk data in more relevant ways to get proper extent of measurement with minimal error percentage. After evaluation of many such methods, listed below are some methods associated with statistical and computer shade:

1) *Analyzing Cluster*: Analyzing cluster is a statistical approach for clustering objects. Specifically, this clustering is done for related objects which share some common features. Cluster analysis is a study method which is not supervised or controlled and hence, not such influence of training data id needed for this case.
2) *Traffic analysis*: This is a method to analyze the incoming and outgoing tremendously huge batches of data for a particular period of time. The analysis is done by reading the information of incoming data and outgoing data, analyzing their role and criteria and finally segregating these data accordingly.
3) *Factor analysis*: This is another type of analytic method concentrating on factors for grouping data. Factors describe closely relating elements and objects. Many variables are termed and data is analyzed and collected according to the preferred batch of variable. These variables then, are grouped on the similarity or relating criteria and termed as factors.
4) *Correlation analysis*: This is another statistical method of analyzing big data in terms determination of correlating parameters for grouping of data. This is quantitative approach of big data handling for control and forecast phenomenon. The data grouping is determined by some uncommon and nonrelating criteria with no dependency. One instance can be stated as a person who bought chips from a supermarket, also bought shoe polish during the purchase.
5) *Regression analysis*: Regression is a mathematical method of analyzing big data. In this tool a particular variable is compared against several other variables, correlations are formulated in between them and so, grouping them in more easier and comprehensive batches. This method can convert complex and huge data into simpler and relevant forms of batches of data.
6) *Bucket testing*: Bucket testing is also termed as A/B testing. This method comprises of several testing and experiments through which big data is processed. Various measures and tests are determined for big data so to check for significant difference to classify control group from treatment group.
7) *Data mining*: This process deals with highly random, fuzzy, noisy and in complete data. Data mining extract meaningful and potential information and knowledge from hidden files. Many databases and decision support software perform the work of data mining.

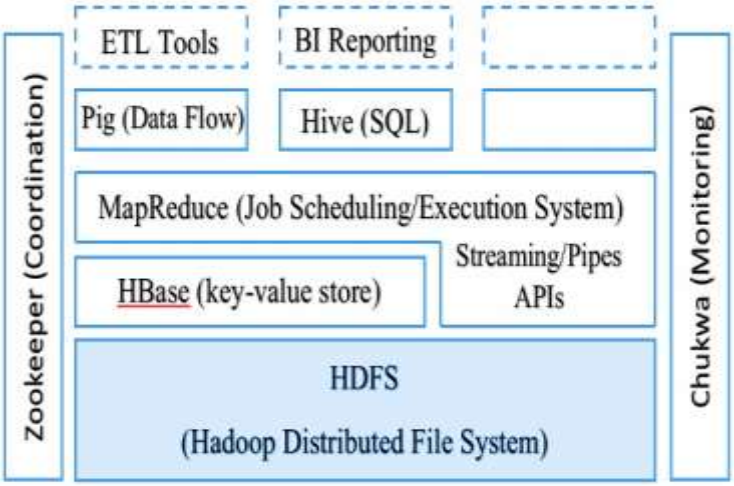
An analyzing approach by Hadoop

[8]: In the year 2005, on the solution of Google, a team with Doug Cutting & Mike Cafarella started an Open Source Project. Later Doug named this project as "Hadoop" after the yellow toy elephant of his son.

Hadoop follows MapReduce algorithm to run different applications. Here processing is done on distinct CPU in parallel manner. Hadoop is efficient to work on applications which are processing on different clusters of computers and thereby producing massive data analysis.

Principles of Hadoop design:

- Self-healing and able to manage: Auto- managerial



  ways in case of failure, as well as execution of reductant tasks in case of slow responses from nodes.
- Linear increment in results and performance: dependent criteria of capacity with difference in resource.
- Movement of computation to data: comparative less in bandwidth as well as latency.
- Facility of extension, modularly operating and core relevancy.

Benefits from Hadoop:

1. Quick processing of massive data: The volume factor of Big Data from IoT to social networking sites, Hadoop shows high efficiency.

2. Power of computing: Hadoop provides the proportional benefits as with high computing more power of processing can be gained.

3. Managerial to failure: Hadoop provides efficient protection against hardware disorder or failure, which can be resulted in nodes' slow operations. Hadoop allows several copies of data to be stored automatically.

4. Durability: Unlike traditional aspects, Hadoop allows to store enormous amount of data according to our desire before deciding the processing.

5. Minimal expense: The open source framework is open for all and storing process is done on commodity hardware.

6. Optimizing size and scaling: With small administrations, Hadoop can made to store and operate more and more data by simply adding more nodes.

Challenges against Hadoop: Firstly,

MapReduce algorithm doesn't provide results for all problems related to Big Data as it fails for iterative and interactive tasks of analyzing. Moreover, being a file-intensive, so the created multiple files because of nodes' interconnection, results in flaws for advanced analytic techniques of computing.

Besides, programmers with MapReduce skills are uneasy to find as Hadoop requires low-level knowledge of OS, hardware and kernel. Apart from this Data insecurity is another limitation, despite surfacing of new technologies and methods. Lastly, Hadoop is not such user interfaced as it is pretty difficult for cleaning data, managing data, metadata and governance.

Overall Evaluation: The amount of data has been increasing and data set analyzing become more competitive than for the previous employed technique. The challenge is not only to collect and manage vast volume and different type of data, but also to extract meaningful value from it. Also needed, managers and analysts with an excellent insight of how big data can be applied. Companies must accelerate employment programs, while making significant investments in the education and training of key personnel.

According to the, Intel IT Center Big Data Analytics survey, there are several challenges for big data: data growth, data infrastructure, data governance/policy, data integration, data velocity, data variety, and data compliance/regulation and data visualization. In addition, Intel IT Center specify obstacles of big data as: security concerns, capital/operational expenses, increased network bottlenecks, shortage of skilled data science professionals, unmanageable data rate, data replication capabilities, lack of compression capabilities, greater network latency and insufficient CPU power. In spite of potential barriers, challenges and obstacles of big data, it has great importance today and in the future.

Conclusion: In this article, an overview of big data's content, scope, samples, methods, advantages and challenges and discusses privacy concern have been reviewed. The results have shown that even if available data, tools and techniques available in the literature, there are many points to be considered, discussed, improved, developed, analyzed, etc. Besides, the critical issue of privacy and security of the big data is the big issue will be discussed more in future. Although this paper clearly has not resolved the entire subject about this substantial topic, hopefully it has provided some useful discussion and a framework for researchers. So, Big Data can be viewed as emergence of developing approaches for modern society and so back-hits and challenges for analytics, researchers and scientists.

Acknowledgement: The authors of this review paper would like to appreciate the encouragement and appreciation by our guide Prof. *Monu Singh* . Moreover, the contributions from various video lectures and conferences on many sites are also in gratitude acceptation. Besides many guest lectures and seminars at our institutional level also enhanced our conceptual knowledge about Big Data. We would convey our regards and thankfulness for all.

References:

1. http://en.wikipedia.org/wiki/Big_data
2. http://www.humanfaceofbigdata.com
3. http://www.zdnet.com/article/top-10-categories-for-big-data-sources-and-miningtechnologies/
4. https://datafloq.com/read/five-dimensions-make-big-data-infographic/264
5. https://upside.tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx
6. https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs/ ; https://www.slideshare.net/HarshMishra3/harsh-big-data-seminar-report
7. https://link.springer.com/chapter/10.1007/978-3-319-06245-7_5
8. https://www.sas.com/en_us/insights/big-data/hadoop.html ; https://www.tutorialspoint.com/hadoop/hadoop_big_data_overview.htm
9. https://www.enterpriseirregulars.com/102734/2015-big-data-market-update/