ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 3, March 2018

Big Data:A New Approch for Data Processing

Vaibhav P Narkhede [1], Sachin Vyawhare [2], Priyanka Kharche [3], Sachin Dandge [4]

Assistant Professor, Dept. of CSE, PLITMS, Buldhana, Maharashtra, India[1]

Assistant Professor, Dept. of CSE, SEC, Washim, Maharashtra, India[2]

Dept. of CSE, PLITMS, Buldhana, Maharashtra India[3]

Assistant Professor, Dept. of CSE, PLITMS, Buldhana, Maharashtra, India[4]

ABSTRACT: With the evolution of large number of technologies and fast growing internet as well as networks, there is a need to capture, process and store the large amount of data created with this means in the market and extract the value from it thus evolved the Big data. Big data analysis consists of analyzing, processing and finding the way to store this massive amount of data for the utilization of organization. It gives a new way of thinking and approaches to analyze and problem solving, as a current research aspect. Based on describing the concepts, characteristics of big data, this paper describes the overall system, its importance and development of technologies, storage, analyses the trends and different values in commercial applications like manufacturing, biomedical science and also many more. At last, we summarize the existing tools and frameworks used for big data analysis and put forward the view that deal with handling the big data challenges correctly.

KEYWORDS: Big Data, Organizations, data analysis etc.

I. INTRODUCTION

We have entered into the era of Big Data which is not just the data but it is more than it. Fundamentally, people need to find a way to extract value from the large amounts of data that pass through their organizations. But big data presents an inherent obstacle because big data is uneven, disparate, incomplete and often in motion. It's hard to get a grasp of big data in a way that delivers value. As society is becoming increasingly more instrumented and as a result, organizations are producing and storing large amounts of data. Managing and gaining insights from the produced data is a challenge and key to competitive advantage. Analytics solutions that mine structured ,unstructured, semi structured , odd-even data are important as they can help firms to gain insights not only from their privately acquired data, but also from large amounts of data publicly available on the Web or from the social media like facebook, twitter etc . The ability to cross-relate private information on consumer preferences and products with information from tweets, blogs, product evaluations, and data from social networks opens a wide range of possibilities for organizations to understand the needs of their customers, predict their wants and demands, and optimize the use of resources.

With the growth of the world, the data also increased and the storage space requirement also increased. Due to this there various challenges and issues comes out that should be handled. The world with growth generates the massive and complex data that defines the basic concept of the Big Data. The computer science goal is to extract the information from the massive and complex data sets by some analysis and technologies and use that information for the decisions. Big data can be defined in several ways just like it is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. We define "Big Data" as the amount of data just beyond technology's capability to store, manage and process efficiently. These imitations like storage, manage, process, analysis of data are only discovered by analyzing the data itself, explicit processing needs, and the capabilities of the tools (hardware, software, and process) used to analyze it. As with any new problem, it is recommended that some new tools must be evolved and can be use for solving the new tasks. As little as 5 years ago, we were only thinking of tens to hundreds of zeta bytes of storage for our personal computers as well as for the organizations. Today, we are thinking in tens to hundreds of terabytes. Real

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 3, March 2018

time applications of big data in different industries like healthcare, network security, market and business, sports, education systems, gaming Industry, telecommunication. Big data technologies process high-variety, high-volume and high-velocity to extract data value and ensure high-veracity of original data.

II. CHARACTERISTICS OF BIG DATA

Volume, Velocity and Variety as the biggest challenges of data management. Big data definition having the following V's properties. These characteristics are very essential in aspect of the analysis of data in the organization for the proper decision making to get the proper results which will be beneficial for getting the profit in the future by avoiding the mistakes or wrong decision taken due to wrong analysis done in the past.

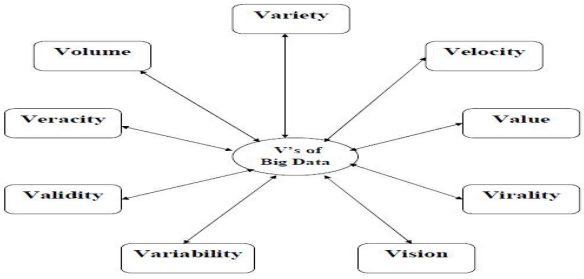

Figure 1: Spider Diagram for Characteristics of Big Data (9V's)

1. Volume: Data volume represent the amount of data available to the organization not necessarily have to own all of it as long as it can access it. As data volume increases, the value of different data records will decrease in proportion to age, type, richness, and quantity among other factors.
2. Velocity: Data velocity measures the speed of data creation, streaming, and aggregation. Ecommerce has rapidly increased the speed and richness of data used for different business transactions.
3. Variety: Data variety is a measure of the richness of the data representation – text, images video, audio, etc. From an analytic perspective, it is probably the biggest obstacle to effectively using large volumes of data. Variety means incompatible data formats, non-aligned data structures, and inconsistent data semantics.
4. Veracity: Veracity means "authenticity with truth or fact" can be simply termed as verification. Data sources are of different qualities with differences accuracy, coverage and timelineness.
5. Value: It refers to the outputs may be midtime in the processing of data. This characteristic of data analysis is very helpful in aspect of decision making for computing. The two
values that can be evaluated by big data are analytical science and data science.
6.Variability : Variability means where the data constantly in changing state. This is particularly the case when data gathering relies on language processing. This is especially true with natural language processing. A single word may have multiple meanings. New meanings are created and old meanings discarded over time specially for synonyms.
7. Vision : Vision is critical in today's world. Using charts and graphs to visualize large amounts of complex data is much more effective in conveying meaning than spreadsheets and
reports chock-full of numbers and formulas.
8.Virality: Virality means the rate at which the data spreads around the globe as it can be used by each everyone for making some analysis for the organization, how often it is picked up and repeated by other users or events.
9. Validity : Validity refers to the time span for which the data is valid for the processing and it is more important for all the researchers to validate the data. The expired data can not been used for the analysis which may results to wrong conclusions.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 3, March 2018

III. SYSTEM FOR BIG DATA ANALYSIS

The system includes number of stages for the analysis of Big Data explained as follows:

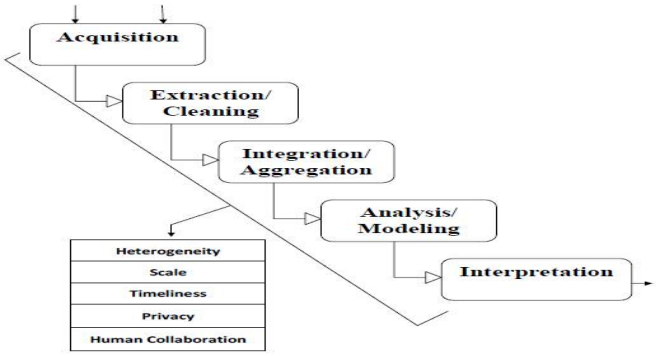

Figure 2: Overall System of Big Data Analysis

1. Data Acquisition and Recording Data can be generated from some sources like social media, internet etc. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information. The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured. This phase only deal with the collection of right meaningful information which can be processed to provide right decision in decision making.

2. Information Extraction and Cleaning At start, the information gathered will not be in a format ready for analysis. Rather we require an information extraction process that gives the desired information from the sources and expresses it in a some of the structured form suitable for analysis. This is one kind of challenge to do this correctly and completely .

3. Data Integration, Aggregation & Representation Due to unstructured data, it is not enough to record it and save it into a repository. Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. This phase deals with the integration of data, its aggregation and its representation in some format before forwarding to the area where it can be save.

4. Query Processing, Data Modeling, and Analysis This phase involves the methods for querying and mining Big Data .All the processing that can be done on the data are given with the queries .Different data format leads to the different models generation which can be helpful for the analysis.

5. Interpretation This phase gives the ability to analyze Big Data is of limited value if users cannot understand the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results.

6. Heterogeneity and Incompleteness When humans process the information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth.

7. Scale At first one can think about the volume of data before the processing it. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades.

8. Timeliness Time is one important factor when we think to process the data. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data.

9. Privacy The privacy of data is another most important concern, and one that increases in the context of Big Data. For electronic health records, there are strict laws governing what can and cannot be done.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 3, March 2018

10. Human Collaboration In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding.

IV. THE IMPORTANCE OF BIG DATA

The main importance of Big Data consists in the potential to improve efficiency in the context of use a large volume of data, of different type. If Big Data is defined properly and used accordingly, organizations can get a better view on their business therefore leading to efficiency in different areas like sales, improving the manufactured product and so forth.In information technology in order to improve security and troubleshooting by analyzing the patterns in the existing logs.In customer service by using information from call centers in order to get the customer pattern and thus enhance customer satisfaction by customizing services.In improving services and products through the use of social media content. By knowing the potential customers preferences the company can modify its product in order to address a larger area of people.In the detection of fraud in the online transactions for any industry. In risk assessment by analyzing information from the transactions on the financial market

Table 1 Traditional data analysis versus big data analysis:

Sr No	Traditional Data Analysis	Big Data Analysis
1	Fundamentally, traditional data analysis contains the analyzing of structured data in predefined format.	In Big Data analysis, there is no restriction on the type of data for analysis .It may be structured, semi structured or poly structured. A real time data can be used for the analysis also.
2	In this type of analysis, normally file based or relational database is used.	In this type of analysis , columnar database is used (Non relational database).
3	Only Static data is analyzed.	Dynamic data is analyzed. It may be real time data or live data in the organization.
4	Traditional methods are based upon the centralized storage of data.	These methods are based upon the distributed storage of data.
5	Basic Knowledge of reporting and analysis tools, few specialized resources.	Advanced analytical, mathematical and statistical knowledge require to develop new models.
6	Uses high cost massively parallel processing computers, utilizing high bandwidth and networks	Innovative methods and virtualized platforms with clusters of commodity hardware.
7	Requires high cost hardware resources to process the data.	It has the ability to deploy the virtualized framework like the hadoop easily to used cloud based environments.

V. BIG DATA FRAMEWORKS & PLATFORMS

The Big Data needs advanced technologies for the purpose of processing of the massive amount of data. These technologies are based upon the analyzing the data, processing the data, manipulating the data and producing the proper output as the decision for the organizations. Following are some of the technologies introduced in the area of Big Data:

1. NoSQL databases There are several types of database that suit into this category, such as key-value stores and document stores, which focus on the storage and retrieval of large volumes of unstructured, semi-structured, or even structured data. They achieve performance gains by doing away with some (or all) of the restrictions traditionally associated with conventional databases, such as read-write consistency, in exchange for scalability and distributed processing.

2. Hadoop Hadoop is by far the most popular implementation of MapReduce, being an entirely open source platform for handling Big Data. It is flexible enough to be able to work with multiple data sources, either aggregating multiple sources of data in order to do large scale processing, or even reading data from a database in order to run processor-intensive machine learning jobs. It has several different applications, but one of the top use cases is for large volumes

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 3, March 2018

of constantly changing data, such as location-based data from weather or traffic sensors, web-based or social media data, or machine-to-machine transactional data.

3. Hive Hive is a "SQL-like" bridge that allows conventional BI applications to run queries against a Hadoop cluster. It was developed originally by Facebook, but has been made open source for some time now, and it's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store. It amplifies the reach of Hadoop, making it more familiar for BI users.

4. Pig PIG is another bridge that tries to bring Hadoop closer to the realities of developers and business users, similar to Hive. Unlike Hive, however, PIG consists of a "Perl-like" language that allows for query execution over data stored on a Hadoop cluster, instead of a "SQL-like" language. PIG was developed by Yahoo!, and, just like Hive, has also been made fully open source.

5. Platfora Perhaps the greatest limitation of Hadoop is that it is a very low-level implementation of MapReduce, requiring extensive developer knowledge to operate. Between preparing, testing and running jobs, a full cycle can take hours, eliminating the interactivity that users enjoyed with conventional databases. PLATFORA is a platform that turns user's queries into Hadoop jobs automatically, thus creating an abstraction layer that anyone can exploit to simplify and organize datasets stored in Hadoop.

6. Wibidata WibiData is a combination of web analytics with Hadoop, being built on top of HBase, which is itself a database layer on top of Hadoop. It allows web sites to better explore and work with their user data, enabling real-time responses to user behavior, such as serving personalized content, recommendations and decisions.

VI. BIG DATA APPLICATIONS

1. Healthcare: Healthcare practices and policies differ tremendously around the world, there are three objectives regarding healthcare system. The first objective is to improve the patient experience (including quality and satisfaction). Second, improving overall population health and reducing the cost of health care and third is traditional methods have fallen short to manage healthcare and create modern technology to analyze large quantities of information. It is time consuming for clinical staff to Collecting massive amounts of data in healthcare. High-performance analytics are new technologies making easier to turn massive amounts of data into relevant and critical insights used to provide better care. Analytics helps to predict negative intervene and reactions.

2. Network Security: Big data is changing the landscape of security technologies. The tremendous role of big data can be seen in network monitoring, forensics. Big data can also create a world where maintaining control over the revelation of our personal information is challenged constantly. Present analytical techniques don’t work well at large scales and end up producing false positives that their efficacy is undermined and enterprises move to cloud architectures and gather much more data, the problem is becoming worse. Big data analytics is an effective solution for processing of large scale information as security is major concern in enterprises. Fraud detection is uses for big data analytics. Phone and credit card companies have conducted large-scale fraud detection for decades. Mainly big data tools are particularly suited to become fundamental for forensics.

3. Market and Business: Big Data is the biggest game-changing opportunity for sales and marketing, since 20 years ago the Internet went main stream, because of the unprecedented array of insights into customer needs and behaviors it makes possible. But many executives who agree that this is true aren’t sure how to make the most of it and they also find themselves faced with overwhelming amounts of data and rapidly changing customer behaviors, organizational complexity and increased competitive pressures. According to Gartner, 50% internet connection between Internet of things (IoT) devices and number reached over 15 billion in 2011 and 30 billion by 2020. Some companies are succeeding at turning that Big Data promise into reality. Those that use Big Data and analytics effectively show profitability and productivity rates that are 5–6% higher than those of their peers. The companies that succeed aren’t the ones who have the most data, but the ones who use it best. Marketing of big data provides a strategic road map for executives who want to clear the chaos and start driving competitive advantage and top line growth.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 3, March 2018

4. Sports: Sport, in business, an increasing volume of information is being collected and captured. Technological advances will fuel exponential growth in this area for the foreseeable future, as athletes are continuously monitored by tools as diverse as sports daily saliva, GPS systems and heart rate monitors tests. These statistics and many more like them are high performance in Big Data. These numbers there is a massive amount of potential insight and intelligence for trainers, administrators, coaches, athletes, sports medics and players. Statistics can be analyzed and collected to better understand what are the critical factors for optimum performance and success, in all facets of elite sport.

5. Education Systems: By using big data analytics in field of education systems, remarkable results can be seen. Data on students online behavior can provide educators with important insights, such as if a student requires more attention, the class understanding of a topic is not clear, or if the course has to be modified. Students are required to answer accompanying questions as they go through the set of online content before class. By tracking the number of students that have completed the online module, the time taken and accuracy of their answers, a lecturer can be better informed of the profile of his students and modify the lesson plan accordingly. The analysis of data also clarify about the interest of student looking at time spent in online textbook, online lectures, notes etc. As result instructor can guide choosing the future path effectively.

6. Gaming industry: The amount of data that video game players are generating on a daily basis is growing rapidaitely. Video game developers are using variety of IT techniques such as Hadoop to keep up the massive amount of gaming data that's generated every day. People are playing video game and generated lot of data in separate areas: game data, player data and session data.

VII.CONCLUSION AND FUTURE SCOPE

As there is rapidly increase in the production of data day to day around the globe we must have to think to preserve this data so that it can be utilized for the processing to get help in making some decisions for the organization. For that many technical challenges described in this must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization at all stages of the analysis pipeline from data acquisition to result interpretation. Furthermore, these challenges will require transformative solutions and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data. We need to provide the solution for this kind of challenges about the big data in future to preserve whole data for the next generation also so that the information will communicate to them which may help them in their future .

REFERENCES

[1] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)
[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.
[3] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
[4] K. Elissa, "Title of paper if known," unpublished.
[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.
[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.