International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

Empowerment through Big Data : Issues & Challenges

Sapna Arora*[1], Manisha Agarwal[2]
[1]Research Scholar, Banasthali Vidyapith, Jaipur, India
[2] Associate Professor, Banasthali Vidyapith, Jaipur, India

ABSTRACT

Big data is a powerful area for narrowing down the useful information from large amount of data which remains unutilized. In a country like India, where majority of populations using devices like mobile phones, laptops and other electronic gadgets, through which an immense amount of heterogenous data generates. This massive amount of data allows different users to work with different types of data that could either be in text, audio or video format in order to fetch out productive information. With a large amount of opportunity to work with, Big data also offers a number of challenges to deal with. Big data has inordinate potential to empower world using proper methods. This study surveys about Big data and the methodologies associated with big data. Also, it gives new insights on how quality information fetching helps in various fields. The study also proposes the different perspectives of big data and future of perceiving data.

Keywords: Web, Big data, Perspectives, Tools, Perceived data

I. INTRODUCTION

The Digital web world offers global users to participate in a rapidly changing online environment, which provides endless opportunities. Each and every user in this web world is indulging themselves through various activities like uploading or downloading any data, streaming of data through an application like YouTube or creating a profile to enter in social media. However, the world seems to be attractive in terms of creating opportunities to work with different data types but also offers a wide range of issues like generation of massive amount of data, capturing & processing of data, heterogeneity of data. In any era, the data or information is one of the invaluable assets for any company or organization. Every organization wants to utilize the information and resources to its maximum. Therefore, valuable information collection, storage and accessing is a major deal which couldn't only be achieved through traditional methods like database which is simple to use. In this information era, the increased use of computers and mobile devices lead to increase in amount of data from few megabytes to Petabytes. It may lead to the use of new tools and storage methods which can handle such situation. Moreover, the perceived data is composed of unstructured and structured types, which is not possible to access through a normal data structure (as it is composed of text, number values like true, false). Formerly, the concept was generated via 3V by Laney [1], according to which to manage explosive data challenges, we must work in three dimensions i.e. volumes, velocity and variety. Later on, in 2012-13, Gartner overhaul this denotation [2] as "Big data is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making." According to him, 3v tends to attract one more factor for analysis process i.e value, which refers to fetch out valuable information from massive data.To be precise, the 10V's are workable acceptable with big data [3]. These are shown in Figure 1

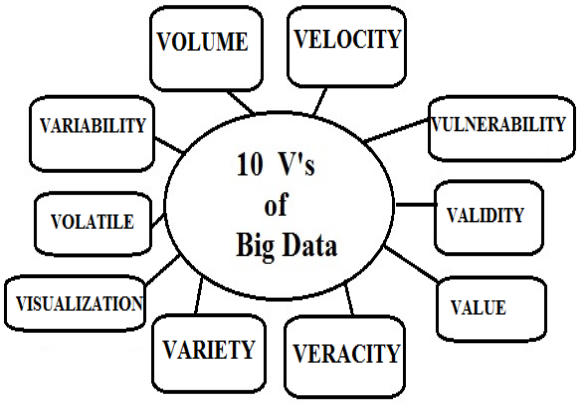

Figure 1. 10 V's of Big Data

A. Volume

The term Volume is meant for the magnitude of data. Massive amounts of data generated from multiple resources are not possible to handle through the traditional ways like a database. This large volume [5] data is a composition of multiple data types, which is unstructured in nature. This kind of data can be either in the form of audio, video, tweets, likes etc.

B. Velocity

Velocity refers to the speed at which the gigantic amount of data is being generated, collected and scrutinized. With every flip of second, data is being searched on the internet. On a day to day basis, social networking sites like Facebook, Twitter, Linkedln etc, are sharing a large amount of data. For easy analysis of this high amount of constantly generating data with keeping an eye on it speed and easy access, it is high time to design and use special algorithms and techniques.

C. Variety

In terms of Big data, term Variety of data pretends to be a composition of structured and unstructured kind of data. The data collected from different sources like mobile phones, laptops etc is not homogenous in nature. Apart from text, audio ,video files, there may be some log files ,clicks or likes or dislikes etc. with which you have to deal with.

D. Value

Value refers to convert our investigated data into values. Value[4] is one of the most important characteristics of Big data with a composition of collection and analysing the same in order to boost the performance of any organization along with a better understanding of customers. With the access to this useful data, one must analyse great values in order to get amazing benefits.

E. Variability

Variability refers to erratic changes in the data. It may happen because of multiple data types & the speed with which data is generating and being loaded into the database. Therefore, it is the high time to develop innovative methods in order to handle the variable data.

F. Veracity

Veracity refers to the term trustworthiness with reference to accurate data. If the data is accurate, only then you could think of meaningful data. For example, consider a dataset of thirty students on which we have to make an analysis about the reason they got distinction. Being an analyser, you can ask questions like: what are the methodology you adopted to get good marks in all the subjects? How much time you devote to individual subject? Do you learn some subjects with the help of daily life activities like sports etc? Have you ever been a scholar? Be getting answers like this it would be easier to determine the accuracy of information which could easily be maintained in statistical form.

G. Validity

Two terms of big data veracity and validity seems to be alike but are quite different. If veracites moves around the management of uncertainty associated with particular types of data then validation is meant for an accurate analysis in order to get optimized results.

H. Vulnerability

Vulnerability is one of the major challenge in Big data as the data generated from multiple sources with such an erratic speed has high chances of being harmed by any intruder. With the increasing use of electronic gadgets and social networking sites, the personal data is on stake. Currently, in a case of facebook, where the Belgium court has threatened to fine a high amount on breaking privacy laws by tracking people on third party websites.

I. Volatility

Volatility refers to how long the perceived data remains to be useful for us and how it is to be kept. For analysing the same, it is necessary to develop some new rules and techniques through which rapid access to information is possible.

J. Visualization

Data Visualization is one of the most complex challenge in big data. In this information age, data is not only going beyond the limits but also is composed of different datatypes. So, there is a need of a way to communicate the information by visualizing it through some special ways with special functionalities like a web-based approach, statistical analysis etc. Traditional tools of data visualization face severe challenges like low response time, complex methods of scalability, precision in reporting time etc. So, it is a challenge to work with the concept which way of communication with data is most suitable in order to make visualization more effective.

II. BIG DATA-DIFFERENT PERSPECTIVES

Big Data is a prominent field that is based on data analysis for decision making. Big data is a combination of complex and diverse set of activities which are explored in various domains like Education, healthcare, management, smart cities etc. Big data has the ability to of procuring empowerment into various domains mentioned below:

A. Big Data in Education

Education plays an important role in empowerment. Famous quote by Nelson Mandela states “Education is the most powerful weapon which you can use to change the world”. With the passage of time, the technology advancements are inclining day by day, because of which every institution has to keep an eye on this upgrading technology with student’s behaviour towards learning. This field could give amazing results if it is combined with big data. Each student is unique has different ability to learn things in a different way. The students who follow a proper way, leads to achieve a good grade but those who don't usually gets a low grade. To improve those low grades, an analysis and availability of a customized program like online courses availability must be able to monitor students effectively and efficiently.

One of the main advantage of inculcating the concept of big data in to education [6] is to be analysing the performance and to get a better result. In this competitive world, every organization or institution needs a good rank holder but what about those who achieved low grades, or the reappear. If such a case is to be analysed, then such a big problem of improving the grades could easily be resolved out. This will help in improving the overall growth of organization.

B. Big Data in Healthcare

Health system in a developing country like India has three tier system[7] i.e. primary level, secondary and medical colleges hospitals. Big data analytics helps in these healthcare sectors by managing quality and timely delivery. In the traditional way, the files of each patient were maintained which got replaced with table-column structure in a relational database management system. With the increase in no. of patients, every year the traditional methods of maintained records are not viable. As the data is

fetched from different sources in and reserved in a distributed file system. Also, this increasing volume of data may give opportunity to find out useful information. Apart from maintaining data, it is also necessary to keep a track of how data is stored and what will be its storage cost. As Big data has a simple designing scenario which is easy to use & because of its ambiguous structure of data, it would be cost effective in comparison to a traditional Database system. In terms of analysis, special technology, tools & technical expertise are required to scrutinize the different effects of big data. Big data is progressively used in health sector to improve services and speed of the processes.

C. Big Data in Smart city management

A city is considered to be a Smart city if it handles most of the real-life problems using technological resources to improve the quality of daily life tasks with optimized costs. Harrison (2010) had given definition of smart city [8] as "A City connecting the physical infrastructure, the social infrastructure, and the business infrastructure to leverage collective intelligence of city". To make a city to be smarter, it is necessary to know that how to control data generated from different ways. Data collection and analysis of large datasets [9,10] in a smart city provide information about the citizens. Using big data analytics, a city can manage some daily life functions. Example: Railway transportation system, where data from a network of sensors that how many persons are in and out at any point of time. Through this data, you could actually analyse the days in which the station is not so crowded, so that you can do necessary activity like repairing of specific area of that station. Apart from it, in a smart city, national census, vehicle transportation and other records could easily be analysed.

D. Big Data and IOT

In this information age, as the number of gadgets users are increasing constantly could result in explosively high amount of data [11] that shows that how data is imbricated with IOT. According to Wikipedia, "The Internet of Things (IoT) is the network of physical devices, vehicles, home appliances and other items embedded with electronics, software, sensors, actuators, and connectivity which enables these objects to connect and exchange data". IOT[12] helps in connecting people (via gadgets) globally using internet. This allows them to access the content anywhere in the world. Main challenge with this kind of data is to implement privacy and to perceive useful information out of it. This can only be achieved using big data analytics. With the inclining quantity of data, the merging of these two technologies leads to a positive change, with the needs of new infrastructures like PaaS, SaaS etc. Only the tools of bigdata have the ability of handling this massive amount of data generated from multiple devices. Big data has the ability to enhance quality & effectiveness of IOT data.

III. EXPLORING THE BIG DATA APPROACH

With the emergence of large amount of Digital data from multiple sources i.e. Social network data like tweets, likes etc, financial transaction-based data from industries, several challenges arise. These can be related to storage, quality, management, analysis etc. The trait associated with this research is to discover the solution for these challenges[17,18]. Formerly, some of these challenges [13] were not able to identify but with the passage of time, the innovative tools and techniques improves performance. In the given section, we rundown some of the tools & Technologies associated with big data.

A. Cloud computing

For the storage of effective & efficient mined valuable information, Cloud computing technology is used. These mining technologies [14,15] helps in fetching valuable information, like prediction of real time traffic operation and safe monitoring using big data, accident prediction, health care etc. But to

manage the data for the same purpose, you need to send it to a specific platform like cloud storage[14], from which a user can share the file with another user. Moreover, cloud storage provides a cost-effective solution to make work cost effective, scaling of policies & plans as per the requirement. Although, there is no need to set up a new infrastructure to work with cloud storage, but special algorithms and techniques are required for data compression for cost management & discovery of data's accurate analytical value.

B. Hadoop

The design and implementation of Big data has to deal with a number of steps in order to perceive the beneficial data.

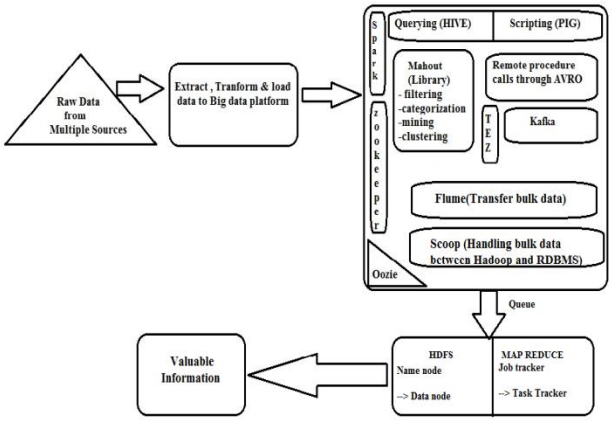

Figure 2

C. Spark

Spark is a developer friendly framework which provides special tools to do various tasks related to big data. These tasks could be related to live streaming data, processing of multiple jobs in batch or large amount of query handling. However, spark is fast and user friendly. platform but it doesn't own its own system for file organization. Therefore, it is used with HDFS.

D. Hive

Hive is a tool that resides in the Hadoop ecosystem. It works by using a language like SQL, which is known as HiveQL[18]. This language helps in handling queries, which are a part of database. This language is compiled using MapReduce. This platform has various subcomponents with it to work efficiently. These subcomponents are Driver, Megastore, execution engine and compiler. Hive platform is dependent upon three fields as tables, buckets and partitions.

E. PIG

Pig[18] is a sub platform of Hadoop ecosystem for analysing massive datasets, that consists of a high-level scripting language, known as pig Latin. Pig supports all types of data i.e. structured, semi structured and unstructured and has its own data type. It enables users to write pig scripts, which are much easier to write and let them convert into a sequence of MapReduce jobs in order to execute through Hadoop clusters.

F. Oozie

Oozie is a component of Hadoop ecosystem (assimilated with MapReduce, Sqoop) which works as a scheduler for different jobs by coordinating, managing & execution at regular interval of time. coordinates, executes, and manages job flow. It works in two parts-one of which is responsible for storing and sunning of jobs, called Workflow engine and other one is responsible for execution of these jobs, is known as coordinator engine.

G. SQOOP

SQL to Hadoop is a Command line interface tool that enables the process to extract data from different data stores of Non-Hadoop structures through the Process of ETL.

H. Mahout

is a library which allow the coders to use a ready to use environment to deal with data mining and machine-learning techniques like Classification, clustering, pattern making etc. This library works well with distributed mode & helps in perceiving of useful information from big data.

I. AVRO

Avro is a component of hadoop ecosystem which works as a data serializer, with specific data

structures to store data. It helps in supporting the process of Remote procedure calls, and schema reducing methods. Avro passes the data from one program type to another. Therefore, integrates well with different types of language like c, c++, C#, Java, Python etc.

J. Kafka

is a fast, reliable, scalable & fault tolerance messaging system which helps in reading data using cluster. Such a messaging and integration system, which helps by entering important messages and topics .

K. MapReduce

is a model which helps in processing massive data by splitting, shuffling, mapping, sorting & reducing. It handles large data through algorithms. Initially, the large data is clustered in to small subsets, which are distributed through clusters . The job tracker interacts with the further nodes i.e. task tracker. MapReduce[11] allocates work to different servers in which the processed information can be kept. Because of its easy access and usage, a company like google is using it in implementation of thousands of programs, which results in processing of big data, which is more than 30 petabytes of data per day.

L. HDFS

Hadoop distributed file system[16] is a reliable and fault tolerant distributed file system that is designed to support large data sets across machines. These machines has the ability to reserve these files in a similar size block, which are arranged in a sequence . HDFS supports parallel processing on data. HDFS works well with MapReduce. Initially data are stacked in HDFS in blocks and distributed to data nodes. Further on, name node interacts with name node, which tracks the blocks and data nodes. HDFS can administer to manage large amount of data and to make different copies of data in order to do parallel processing on data.

Table 1

Hadoop components	Function
SQOOP	CUI to extract data from different data stores
OOZIE	Job scheduler and workflow management
MAHOUT	Library for data mining and machine learning
AVRO	Data serialization
HIVE	Query Handling using HiveQL
PIG	Script writing using Pig Latin language
KAFKA	Instant messaging and integration system
MapReduce	Programming model work on shuffle, map and reduce process.

IV. OPPORTUNITY AND CHALLENGES

Big data is one of the popular area to work for. However, to organize the content is a challenge for every organization. The varying size of heterogenous data [17] makes it difficult for organization to work upon it. Apart from it, to explore data in an efficient way is also a big challenge. To handle big data [18], special data processing tools are needed to fetch out knowledge from gigantic amount of data. However, for minimizing hardware requirement, optimizing cost and efficient processing of this data requires some more technologies to work with. This section reviews the technical opportunities and challenges associated in order to explore this data in an efficient manner.

A. Helix

A system that helps in analysing any type of data using a special mechanism. This system allows data

scientists to work with this massive data with the concept of abstraction [19]. Here, abstraction of data term refers for a data scientist that don't need to emphasize on the outline of data. However, he/she 'll consider the type of data and its usage. Helix users i.e. data scientist, knowledge workers or an expert/s traverse through massive sources of data by using keywords and their combination. With this, they can be able to navigate through different webpages that include information about the basic outline of data and other important details about data, that is released from multiple sources [19,20]. This System is composed of data source registry, data processing unit, query builder, navigational tracker, manipulator and interactive UI. To work with heterogenous data types is a big challenge in big data, which is resolved with helix by binding various metadata from diverse data models that can be a resource description file, audio, video or an XML file. Such a system with common semantic model could make big data related work more cheaper and help in capitalizing more profits from valuable information perceived through the massive data.

B. Sandbox

is a platform which helps in exploring the valuable information from complex analytical data. This data can be a structured, unstructured and streaming data. Various key components of sandbox like business analytics, PCs, Parallel CPUs, BI tools, cloud storage etc makes the work easier by providing special functionalities. Sandbox, also known as spreadmarts [20,22] gives a support to deal with constantly changing data, that is perceived through a number of sources. By using sandbox, organizations can reach to efficient decision making. For an IT professional, Data Scientist or a researcher, a perfectly created sandbox provides a safest place to deal with massive data of organization, which is a challenge in big data. By using the concept of sandbox, one could achieve the richest information sets from the big data.

C. Pentaho

A high-performance platform which allows users to create reports [21] using heterogenous types of data. These reports can be generated in any format i.e. .RDF, .pdf, .XML, .html etc. Data can be extracted from multiple data sources, integrated & converted into reports of desired type. It is one of the open source product which is available to use for data integration, security, report outcome, scheduling of tasks & analysis with visual tools that make it easy to define the data for effective analysis, which is a big challenge in big data. These data sets could easily be apportioned with another user. Most important fact with this technology is to understand the basic concepts associated.

Though, these platforms provide efficient processing, yet challenges in big data like security, privacy needs special attention. Also, future research needs to deal with privacy issues related to big data. Apart from the privacy, data integrity also needs to be emphasized. As the loss of data could lead to a great loss for organization. Therefore, more research is needed to face these issues in order to make effective analysis, storage and privacy of big data.

V. FUTURE OF PERCEIVING DATA

These future research work initiatives could make the strong fortunes of empowerment in every domain. With the passage of time, more and more data will be exposed to the web. To make smart and efficient data processing methods, a lot of fact finding is needed with more precise analytical methods. Lot of work must be done on data compression while dealing with big data. Using advanced compression techniques, one could utilize the space very efficiently. Most of the cloud services provide limited bandwidth, which may affect on the budget of an organization. Some more opportunities may give an insight view of a new concept like some valuable information available in old data, which is being overlooked by new data. Special Techniques &

Algorithms need to be developed to recover these valuable information & effective visualizations of massive data.

VI. CONCLUSION

The paper describes the concept of 10Vs which are the main issues before a data scientist to deal with. The paper also emphasizes on the challenges and problems with the big data. In this study, we presented a review on the empowerment of big data in every domain that is related to our real life, with different perspectives. The review also summarizes the tools and techniques used for addressing the big data processing problems. The review also enlightened several challenges faced with big data processing, gives a chance to number of opportunities to work with new technologies like helix, sandbox etc.

VII. ACKNOWLEDGEMENT

We would like to thank Mrs. Shweta Mongia from whom we took help during writing of paper.

VIII. REFERENCES

[1]. Storey C., Song, Yeol-II. 2017 Big data technologies and management: What conceptual modelling can do., Data & knowledge engineering, vol.108 pp.50-67, Elsevier.

[2]. Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s: https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/#18e54f8d42f6

[3]. https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

[4]. Yang, Chaowei, Huang, Qunying, Zhenlong Liu, Kai,Hu Fei.2016. Big data and cloud computing : innovation, opportunites and challenges, pp:13-53, Taylor and Francis Group

[5]. Jayant Madhavan, Afanasiev Loredana, antova Lyublena; Halevy, Alon.2009. "Harnessing the deep web: Present and future", CIDR

[6]. Thakkar Pooja, Mehta, Anil, Manisha. 2015. Performance analysis and prediction in Educational data mining: A research travelogue, vol.110, International Journal of Computer Applications

[7]. Garg Ravish, Singh Atul .2010." Proposed design for healthcare system in Remote Areas" In Proceedings of National conference on Role and Application of Information and communication technology in inaccessible areas, pp 263-265.

[8]. Saraju P Mohanty, Choppali Uma, Kougianos.2016. "Everything you wanted to know about smart cities: The Internet of things is the backbone", pp.60-70 IEEE

[9]. Chen M. 2014. Related Technologies in Big Data. Heidelberg, Germany: pp. 11-18, Springer

[10]. Kitchein Rob. 2014. The real-time city? Big data and smart urbanism, Geo Journal, pp.1-14, Springer

[11]. Dean Jeffrey and Ghemawat Sanjay.2008. MapReduce: Simplified Data Processing on Large Clusters, In Proceedings of the 1997 ACM SIGMOD International Conference, pp.1-13, ACM

[12]. Steinklauber.K. .2014. Data Protection in the Internet of Things. Online: https://securityintelligence.com/data-protection-inthe-internet-of-things

[13]. Li, H., Lu, X. 2014 "Challenges and trends of big data analytics." In: Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, China. pp. 566-567.

[14]. Ji Changqing, Li Yu, Qiu Wenming.2012. Big Data Processing in Cloud Computing Environments, IEEE

[15]. Burtica R., E. M. Mocanu, M. I. Andreica, and N. Ţăpuş.2012. Practical Application and Evaluation of No-SQL Databases in Cloud

Computing. In Proceedings of IEEE International Systems Conference (Sysco), pp: 1-6.

[16]. Shvachko K., Hairong, S. Radia, R. Chansler.2010. The Hadoop Distributed File System, Mass Storage Systems and Technologies (MSST), IEEE 26th Symposium on, pp. 1-10, ACM

[17]. Data Modelling and Data Analytics. 2015.: A Survey from a Big Data Perspective, Journal of Software Engineering and Applications, pp. 617-634 online: https://file.scirp.org/pdf/JSEA_20151230145049 42.pdf

[18]. Orgaza Gemabellow, Jason, Camachoa David. 2015. Social big data: Recent achievements and new challenges, Information fusion,pp:1-14.

[19]. Chen X.Y. and Jin Z.G. 2012. Research on key technology and applications for internet of things, Physics Procedia, 33, pp. 561-566 .

[20]. Lipic Tomislav, skala Karolj, Afgan Enis, "Deciphering Big data stacks: an overview of Big data tools"

[21]. Ahmed Abbasi, suprateek sarker, Roger HL chiang,2016. "Big data research in information system: towards an inclusive research agenda" ,Journal of the association for Information systems 17(20),

[22]. Sherman Rick.2013. "Maximize your ROI from business intelligence" Analytics Best Practices: The analytical sandbox: white paper.