Big Data Challenges and Prospective Solution

[1]Mujtaba Ashraf Qureshi, [2]Dr. Azad Kumar Shrivastava

[1]Scholar, Deptt.: Information Technology, Mewar University, Gangrar, Rajasthan.

[2]Professor, Mewar University, Gangrar, Chittorgarh, Rajasthan

Abstract: Data is powerful raw material and acts a baseline for numerous research fields and for other related services. We are approaching to the age of big data which consists of very much diversified datasets that are characterized by complexity, high volume, redundancy, diversity etc. The derived knowledge from these huge and diversified datasets plays an important role to make strong decisions and conclusions in future and present. However, to deal with such datasets is not possible by means of traditional methods, thus it has become a problem. Thus it requires a very special form of soft wares or platforms. In this research work different challenges faced by big data and its prospective solutions are discussed in detail. Here also a discussion related to framework, platforms with respect to mining big data. Also various tools and technologies related to big data are discussed. Finally the issues, current efforts and future work related to big data mining presented.

Keywords: *Big Data, Challenges, Tools, Issues, Future scop*

I. INTRODUCTION

There is no any doubt that we are living in digital world today, as almost everything exists in digital formats. Data is generated everywhere and every second and micro second data is generated at alarming rate and gets stored at different locations. Thus it is more suitable to call big data era than data era to present world. To process complex and redundant data is very difficult by the applications of general traditional and conventional tools i.e. traditional tools are enough to deal with such enormous and big data. The big data is existing in number of different format le video, audio, text etcetera (structured, semi structured & unstructured). Also the size of data varies from petabytes to zeta bytes or more than that. There are many problems existing related to big data and the primary problem in big data is the lack of compatibility existing between different databases and tools of analysis. There appears challenges in big data analysis at the time of discovery of knowledge and representation from big data. Let's see now what the big data exactly is.

[1]*Big data*

There are plentiful and abundant sets of data available in diverse locations to draw some useful knowledge, trends, patterns and associations are called big data. Big data refers to large volumes of structured and unstructured data existing in different databases. In big data quality is considered more significant than quantity of data. There are many sequential steps adopted in this process like data gathering, processing and analyzing of data but final aim lies to acquire only useful patterns, valuables and knowledge from that big data. If the knowledge hidden in the big data is done precisely and exactly better and improved decisions are inevitable for humankind. Big data is defined under three headings by data analyst as; ***Velocity:*** Data is increasing at enormous speed from different sources and gets stored under diverse databases. Due to increase of data at alarming pace needs specialized mechanisms and tools to handle. ***Volume:*** Data is collected from different platforms such as business organizations, health systems, social sites, transactional data and much more. ***Hadoop:*** This is proved a successful and really good at managing very higher contents of volume of data.

1.2 Purpose of this Research Work:

An exploration related to challenges and other open issues with respect to big data is presented. An open discussion regarding technologies and tools employed to deal with big data challenges is also presented. An investigation of the effects and influence of big data in a variety of domains is performed. In this paper some of the prominent issues are put forth to help research in big data at a good level.

1.3. Paper Organization

This research paper has been well organized into different sections as; section 1 presents an introduction and challenges to big data. Section 2 defines big analysis and preprocessing issues. Section 3 clarifies the solutions for big data challenges. Section 4 defines the ways and methods to be followed to process and acquire useful patterns and knowledge. Section 5 gives a brief introduction regarding tools and technologies employed for big data. In section 6 the scope of future work is presented. At the end in section 7 a brief conclusion and summary is made available.

II. CHALLENGES OF BIG DATA

It is well known proverb that "after difficulties and challenges there is success". Opportunities always are linked with some kind of hardships so to get them needs a dedicated and well planned platform. Similarly to acquire the insights and knowledge from big data is full of challenges. So suppress or fight with these challenges there is need to know various threats related to security, complex nature of data, computational complications and computational techniques of big data to examine big data difficulties. Let us know that existing methods of mathematics and statistics are well suited for small data sets and find very less efficient for larger data sets. In similar manner various existing methods and techniques are inefficient in computational power to deal with messy and big data. There are many challenges included in primary problems; such as large and abundant volume of data, different structural and semi structural data, variety, to combine different sets, high velocity, data extensiveness, and quality of data, data processing and management. Some of the following noteworthy challenges are presented to define various challenges of big data.

Volume: Size of data is in explosion state. Every year it shows geometric progression in size and volume. It is concluded that in very near future we have a collection of petabytes of data or zeta bytes everywhere. There is very much increase in data via social websites using mobile phones in every Nano seconds [2]. Thus it has become a very difficult challenge to tackle such a messy and large volumes of data.

Velocity*:* Data is increasing at very high velocity from diverse industries and organizations. Now the question arises here how to handle such overflow of data when there exist any special technologies to handle it [3].

Quality and Significance: The dataset collected and processed must be very relevant to the problem and if they were as per the problem, it would be very havoc for the organization. So to determine the quality and relevance of datasets is a challenge [4]. If there is supervised or unsupervised problem in existence so carefulness is need of the hour to select data as per the problem.

Privacy and Security: To maintain the privacy and security of a citizen is the primary role a data analyst or data scientist. There should not be any compromise about the data privacy policies of people. Instead a proper consent must be taken from organization or people whose data we are going to employ. To acquire knowledge and find trends, no care is provided about the security and privacy of a person [5]

Scalability *[6, 5]:* As there is good discussion about the big data increment at alarming pace. In other words we can say that the quantity of data is increasing and gets a stored in diverse databases in such a manner that to scale such volumes has become a very tedious task. To handle such unlimited data for the process of scaling and in single application with fastest speed is very difficult. So there must be steps taken to handle such a problem of scaling of data so that it may not become a problem in future.

In addition of the above mentioned challenges in the field of big data mining some more challenges are given ahead. To detect problems linked with ecosystem, problems related to space agencies, distributed storage, content validation, to find exact relationships etc. Data analytics challenges [7] include: Data storing and analysis, Knowledge discovery problem, Computational difficulties, Scalability and visualization of used data and information security and privacy.

III. BIG DATA CHALLENGES – EXISTING SOLUTIONS

3.1 Data Volume Challenge – A Potential Solution [8]

Hadoop: This is an open source apache technology in which a network of many computers is used to solve the problem of large amounts of data. This software has great capability to tackle the processing of big data in spite of distributed data also. There must an approach to train the related professional so to deal with big data issues and thus to find related insights and knowledge from the data.

Robust Hardware: This is technology which solves many problems such as fault management to help to maintain the integrity or to avoid crashing or failure of a system.

Grid Computing: This is interconnection of large and high processing computers and forms a virtual supercomputer to perform complex tasks such as to tackle the problem related to big data processing, analyzing etc. Thus it provides high storage availability and high processing power capabilities for big data.

Visualization: The visualization techniques are most acceptable techniques to translate large volume of data sets and metrics into graphs, bar charts, boxplots and other visuals to and thus to get good insights from data. Thus it has increased chances to get real knowledge and real trends from the underlying big data.

Spark: This platform uses model plus in-memory computing to create huge performance to gain from large volumes and different databases.

Such type of approaches provides great relief to different organizations and industries to get required insights and knowledge from the data. From the discussion performed above shows that to deal with big data could be performed either to shrink data or to invest huge amounts on the development of efficient platforms to solve big data problems.

3.2 Data Variety Problems- Potential Solutions [9]

There are many existing solutions to data velocity problems and some of the as the use of ; *On-line Analytical Processing Tools (OLAP*) which makes data available in lucid and logical manner. OLAP processes all formats of data provided to them in spite of their variety. *Hadoop* has also ability to tackle the problems of different data formats. As it consists of different computers connected with each other. SAP HANA is a real time in-memory application suitable to perform real-time analytics and deploying real-time applications. It swiftly makes identification of different data formats.

3.3 Data Velocity Problems- Potential Solutions

Some of the prominent techniques, arrangements and technologies employed to solve data velocity problems are transactional databases, flash memory, hybrid model in cloud computing, SAAS,PAAS,LAAS and other sampling technologies. All these technologies have been proved very efficient to solve data velocity problems in big data.

3.4 Data Quality and significance [10, 11]

Data Visualization and algorithms are two methods employed to enhance the quality of big data employed datasets. Visualization presents clustering of related data, explanation using axis methods etc. Different algorithms and techniques are existing to deal with enhancement of data quality.

3.5 Potential Solution for Privacy and security

Some of the prospective solutions to maintain the privacy and security of data are, must have adequate control over data and databases via special policy control policies, examination of cloud providing facilitators, best secure communication paths, authentication methods, real time security methods, logging methods etc.

3.6 Scalability – A proposed Solution

Cloud Computing is the solution for scalability in much easier and faster manner in comparison to other. Keeping data protected in a cloud can solve half a problem because the cloud can be secured and cloud can be extremely spread out that we can call it closely limitless.

IV. OPEN ISSUES IN BIG DATA

The issues associated with big data, if resolved, would be proved very much effective for a country in various fields such as to boost economy, health system, education system, travel industry etc. The studies of various resources and articles have shown some of the prominent to mention as open issues in big data. Some of these most prominent and noteworthy open issues existing in big data are shown as follows;

- Finding relevant data
- Integration Problem of machine learning.
- Better Need of Data Management and Data Analytics
- Need of Better Secure Cloud Platforms.
- Security and privacy of Data.
- Big Data Storage Issues
- Data Management Issues

V. BIG DATA TOOLS AND TECHNOLOGIES

There are various existing tools and technologies utilized in field of big data. Some of them are depicted below;

Hive: A data warehousing application that delivers an SQL-like admittance and relational model.

Sqoop: A project for moving data between relational databases and Hadoop.

HDFS: A highly faults tolerant dispersed file system that is responsible for storing data on the clusters.

MapReduce: It is a parallel programming method for distributed dispensation of huge quantity of data on clusters.

HBase: A column-oriented disseminated NoSQL database for arbitrary read and write access.

Pig: This is high-level data programming language for examining data of Hadoop computation.

Oozie: An instrumentation and workflow organization for reliant on Hadoop jobs.

Hadoop and MapReduce: Hadoop and MapReduce are cooperative in fault-tolerant storage and high output data handling. This tool performs processing at very high rate and works on the principle of divide and conquer algorithm.

Apache Mahout: In its core there are machine learning techniques to resolve big data issues to a certain level efficiently. Some of the techniques existing in Mahout are clustering, classification, regression etc. It is one of the smart platforms.

Spark: This framework has specility of data processing and analytics at high rate. This platform is used to develop program in java, scala, or python. Incorporation of spark applications in hadoop is very much beneficial.

Storm: It is free and open source, devised for real-time processing but not for batch processing. It shows great performance when employed.

Apache Drill: It is a distributed system for communicating examination of big data. It supports many query languages, several data formats. It is exactly for misusing nested data. It has an astonishing ability to scale up to thousands of servers or even more and has the competence to process massive data and trillions of records in seconds.

Splunk: It is a real-time and best platform shaped for creation of the best use of machine-generated big data. It combines the up-to-the-moment cloud skills and big data very quickly and effortlessly.

VI. FUTURE SCOPE OF BIG DATA

Data is generated in large volumes with the passage of every nanoseconds of time in this modern and digital world. So better instructions and resources could be employed to develop new methods and technologies to tackle different issues and challenges in the field of big data. If different organization whether private sector or public sector could join hands to utilize the knowledge of big data there would be billions or trillions of dollars back as revenue. So we can say that big data knowledge has tremendous capability to modernize every organization or work culture by developing well suited future decisions. There are enough opportunities to develop different kinds of soft wares and hardware for present and upcoming requirement in big data.

So we can say that there is very high demand of the services and technologies which could be proved helpful to tackle the issues and problems of big data. The primary problem which needs a very good attention is the storage of large and unbeaten volumes of data with proper security and integrity. Thus steps could be taken to strengthen such existing weaknesses in big data. No doubt there are some existing technologies like hadoop, but they also needs further up gradation and innovation and thus to make them intelligent more. Also technologies related to scalability, quality control, variety etcetera needs to upgrade at present and also enable them for future applicability. Data visualization tools more demanded by big data scientists as they are proven more helpful to get proper insights from large sets of data in more efficient manner. So better visualizations methods can be established better than existing tools. Data security and data privacy has become an issue in the field of big data so steps could be taken to strengthen data communication policies such as user login approach, authentication, data management etc.

VII. CONCLUSION

In this research paper possible problems and challenges existing in big data are highlighted and acceptable proposed solutions are also presented. Open issues related to field of data mining are also discussed. Finally future work in the field of data mining is presented. This research paper will be proved a tool to guide upcoming researchers and scholars.

VIII. REFERENCES

[1]. http://www.sas.com/en_us/insights/big-data/what-is-big-data.html as of February 2016

[2]. VijayaBaskaran, R. "AN ANALYSIS OF EMERGING TRENDS IN BIG DATA AND DISCRETIONARY OPPORTUNITIES FOR INDIAN BPO INDUSTRY." *International Journal of Information Technology & Computer Sciences Perspectives* 2.2 (2013): 441.

[3]. Tole, Alexandru Adrian. "Big data challenges." *Database Systems Journal* 4.3 (2013): 31-40.

[4]. Ammu, Nrusimham, and Mohd Irfanuddin. "Big Data Challenges." *International Journal of Advanced Trends in Computer Science and Engineering* 2.1 (2013): 613-615.

[5]. Nasser, T., and R. S. Tariq. "Big data challenges." *J Comput Eng Inf Technol 4: 3. doi: http://dx. doi. org/10.4172/2324* 9307 (2015): 2.

[6]. Guillermo Lafuente, "*Big Data Security - Challenges & Solutions*" available at https://www.mwrinfosecurity.com/our-thinking/big-data-security-challenges-and-solutions/ **as of** 10 November 2014

[7]. Debi Prasanna Acharya, Kauser Ahmed P, "A Survey on Big Data Analytics: Challenges, Open
Research issues and tools" available a at https:/www.researchgate.net/pubilations/296550027,2016

[8]. Zicari, Roberto V. "Big data: Challenges and opportunities." *Big data computing* (2014): 564.

[9]. Tole, Alexandru Adrian. "Big data challenges." *Database Syst J* 4.3 (2013): 31-40.

[10]. Keim, Daniel A., et al. "Challenges in visual data analysis." *Tenth International Conference on Information Visualisation (IV'06)*. IEEE, 2006.

[11]. Prakash Janakiraman, Co-Founder, and VP Engineering *Nextdoor* "*Big Data Challenges*" Available at https://www.qubole.com/resources/solution/big-data-challenges/#sthash.hls6Z7N7.dpuf as of 11 February 2017