A Review On Big Data Issues And Challenges

[1]Kunal Choudhary, [2]Ishika Gupta, [3]Manoj Diwakar, [4] Ms Manika Manwal

[1,2,3]Graphic Era deemed to be University Dehradun, Uttarakhand, India.

[4] Assistant Professor, Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun.

ABSTRACT

This article presents an overview of big data as well as an investigation into the challenges and problems that are currently connected with the application of big data. Data is being produced at a rate that is exponentially increasing, and there will be an extremely large amount of data produced in the years to come. As an illustration, Facebook generates about 4 petabytes (PB) of data per day, and this rate is only expected to increase in the future. Our conversation will center on the most significant problems and obstacles we face while dealing with this data.

Key words: Big data, Massive Data, Structured big data.

1. INTRODUCTION

The term "big data" refers to data collections that are extraordinarily extensive. The concept of "big data" refers to a massive volume of data that is continually being added to and expanding at a rapid rate [1]. Big data analytics can utilize data from a variety of sources, including social networking websites, jet engines, and financial markets. The major categories of big data are given below:

a) Structured big data:

Structured data is any data that can be accessed, processed, and stored in a predetermined format. Computer science has, throughout time, been increasingly effective in inventing techniques to manage and extract value from this kind of data (assuming that the format is completely known in advance). We're bracing ourselves for problems in the near future as the volume of this data grows exponentially; average sizes are now in the zettabytes range.

b) Unstructured big data:

The challenges presented by unstructured data are numerous and extensive, and they need to be overcome if any value is to be derived from the massive amount of this type of data that exists. Data sources that include simple text files in addition to photographs, videos, and other sorts of data are examples of heterogeneous data sources. Unstructured data is typically found in these types of data sources. In today's business world, organizations have access to vast amounts of data; but, because this data is in its raw or unstructured form, these organizations are unable to derive value from this data.

c) Semi-structured big data:

Data that is semi-structured has both structured and unstructured information in its make-up. Even though semi-structured data may give the appearance of being structured, it is not defined by the concept of a table that is used in relational database management systems (RDBMS). XML files are an example of semi-structured data. These files store data.

2. Big Data Issues and Challenges

i. **Lack of knowledgeable Professionals:** Businesses need knowledgeable data specialists in order to effectively manage the cutting-edge technology and massive data tools at their disposal. To assist in making sense of the massive amounts of new information that will be generated as a result of the implementation of these new technologies, a large number of data scientists, analysts, and engineers will be employed. The inability to locate skilled workers with experience in big data is one of the challenges that any company must overcome. Even though the vast majority of working professionals are unaware of this progress, the technologies used for data processing are undergoing rapid improvement on a regular basis. In order to narrow this gap, actions that produce results need to be conducted.

ii. **Lack of proper understanding of Massive Data:** Businesses that fail to appropriately exploit big data are doomed to fail due to a lack of expertise. Employees may not be familiar with the meaning of the term "data," as well as its sources, processing, or storage. Even while data experts understand what is going on, it is possible that others do not. It is possible that employees will not preserve a backup of sensitive content if they are unable to comprehend the significance of the storage system. They may be unable to maintain data in databases in an effective manner. It is challenging to get such data at the times when it is required.

iii. **Data Growth Issues:** Finding a suitable storage solution for massive amounts of data is among the most difficult challenges that come along with working with large amounts of data. The amount of data that is being stored in data centers and company databases is growing at an alarming rate. The management of big data sets becomes more difficult to accomplish as their sizes continue to grow. It is easy to become disoriented when faced with a large number of documents and other types of media. The fact that they are not included in the database is an encouraging indicator.

iv. **Confusion while Big Data Tool selection:** When confronted with tremendous responsibilities, companies often find it difficult to select the equipment that is the easiest for processing and storing information respectively. Comparing HBase with Cassandra, which one is simpler to work with? How significantly more effective is Spark in comparison to Hadoop MapReduce when it comes to the storage and analysis of data? Despite the fact that firms are looking for answers to these issues, it might be challenging to locate them. They have come to realize that majority of the decisions they make are poor, and that they choose ineffective technologies to work with. As a direct consequence of this, resources such as cash, time, effort, and working hours are thrown away.

v. **Integrating Data from a Spread of Sources:** A company will gather information from numerous sources, such as websites, enterprise resource planning (ERP) software, customer records, financial reports, emails, PowerPoint presentations, etc. It could be challenging to put up a report with all of this information. Businesses have a tendency to miss this location rather frequently. Because of this, the utilization of data integration is appropriate within the parameters of this setting.

vi. **Securing Data:** One of the most challenging issues that big data puts in front of us is to ensure the security of these enormous data stores. Because they are so preoccupied with processing, storing, and analyzing their data sets, many companies fail to recognize the significance of maintaining data security. It is in everyone's best interest to avoid leaving data repositories vulnerable if they wish to prevent inviting malicious hackers. Businesses risk incurring losses of up to $3.7 million in the event of a breach of sensitive information, such as documents or knowledge.

3. Big Data Analysis and discussions:

The best and the brightest people are becoming increasingly difficult to find and keep, so businesses are stepping up their recruitment and retention efforts. Educational opportunities need to be made available to current workers if the company wants to get the most out of the work they are already putting in. One further significant step that companies take is to purchase information analytics solutions that are powered by artificial intelligence and machine learning. Big Data Tools are commonly advocated by data experts that only have a fundamental understanding of data science. When it comes to the process of hiring new employees, this might end up saving businesses a substantial amount of money.

Training and educational opportunities related to Big Data should be made available to all staff members. Every member of staff who routinely works with huge volumes of data or who contributes to large-scale data initiatives is expected to undergo training similar to that received in the military. The entirety of the organization, including all of its employees, must demonstrate a fundamental understanding of ideas such as information and knowledge.

Computing on the cloud and storage that is defined by the software are two approaches that can be helpful in managing the ever-increasing volume of data. By implementing strategies like compression, tiering, and deduplication, it is possible to cut down on the expenses

associated with data storage. These technologies include Hadoop and NoSQL databases, Spark, artificial intelligence (AI), machine learning, and business intelligence (BI).

Either that or a group of professionals who are well-versed with these equipments will be brought in to help. One more alternative is to approach a company that specializes in providing consultancy services for large amounts of data. In this section, the advisor will lay out the solutions he or she recommends for the problems that currently exist inside your firm. After discussing the matter with them, a plan of action will be developed, and the most appropriate tool will be selected.

It is necessary to acquire the appropriate technology in order to cope with data integration. Businesses are increasing the number of cybersecurity professionals they recruit in order to improve the security of the data they hold. The following are also included in the security of large amounts of data: the safeguarding of the personal information of each individual and the maintenance of their confidentiality and the distinction between the various tasks and duties involved in data processing jobs. Within the realm of the internet, both access and identification are subject to stringent regulations. The steps that must be taken in order to obtain a secure endpoint are Real-time surveillance and monitoring of the security situation and utilising the enterprise-level data protection technologies provided by IBM Guardian.

4. Conclusion

Big data and the technologies that are related to it sparked our interest, so we decided to look into the characteristics and limitations of space data and explore them within the context of big data. We devised workable answers to the challenges and roadblocks that we encountered. In today's world, having a competitive edge in a data-driven economy is absolutely vital. It is essential to remember, as a project moves forward, that each individual tackles the difficulties presented by big data in their own unique way. The ever-evolving nature of Big Data can be attributed to the expansive nature of its breadth. Even the most seasoned professionals are capable of developing original solutions when presented with the challenges posed by Big Data.

REFERENCES

[1] For Big Data Analytics There"s No Such Thing as Too Big The Compelling Economics and Technology of Big Data Computing March 2012 By: 4syth.com Emerging big data thought leaders

[2] http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data

[3] Big Data, Bigger Opportunities – Data it. gov"s roles: Promote, lead, contribute, and collaborate in the era of big data ,Jean Yan ,April 9, 2013, www.meritalk.com/pdfs/bdx/bdx-whitepaper-090413.pdf

[4]   http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html

[5] Understanding Big Data: A Management Study, September 22, 2011

[6] Mark Troester(2013), "Big Data Meets Big Data Analytics", www.sas.com/resources/.../WR46345.pdf

[7] http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[8] Divyakant Agrawal, UC Santa Barba, Philip Bernstein, Microsoft Elisa Bertino,...(2012), " Challenges and Opportunities with Big Data",

[9] http://en.wikipedia.org/wiki/Big_data

[10] http://blogs.wsj.com/experts/2014/03/26/six-challenges-of-big-data/

[11] http://cacm.acm.org/magazines/2014/7/176204-big-data-and-its-technical-challenges/abstract

[12] http://www.sas.com/resources/asset/five-big-data-challenges-article.pdf

[13] http://www.datastax.com/big-data-challenges nessi white paper, December 2012,big data, new world of opportunities.

[14] Shilpa et al., Big data and Methodology: A review, International Journal of Advanced Research in Computer Science and Software Engineering 3(10), October - 2013, pp. 991-995