International Journal of Research in Engineering, Science and Management
Volume 5, Issue 6, June 2022
https://www.ijresm.com | ISSN (Online): 2581-5792

A Survey on Big Data Analytics

Shashank S. Naik[1], Divya S. Naik[2*]
[1,2]Student, Department of Computer Science and Engineering, Ramaiah Institute of Technology, Bengaluru, India

***Abstract*: Each day, new data systems and new technologies such as the cloud computing and IoT (Internet of Things) generate a huge reservoir of information (facts and figures) in terabytes. These huge information sets require hard work at various levels to extract relevant information for arriving at a conclusion. Hence analyzing big data is therefore the main topic for research. This study's main objective is to look into the potential impact of big data challenges, open research questions, and related tools. Because of this, this page can be used as a beginning point for research into big data at different levels.**

***Keywords*: Big data, structured, unstructured, semi-structured data.**

1. Introduction

In the digital world, data is produced from many different sources, and big data has been produced as a result of the quick uptake of digital technology. It enables evolutionary advancements in a variety of fields with the accumulation of enormous information. It describes a collection of complicated large amounts of data that are challenging for processing utilising standard software.

These come in petabyte-scale and larger sizes. It falls between 3 and 4 volts. The three Vs are volume, velocity, and variety. Volume describes the enormous amount of data that is produced every day, whereas Velocity describes the rate of increase and the speed at which the data is obtained for analysis. Structured, unstructured, and semi-structured data are only a few examples of the various types of data that exist. The fourth V is veracity, which also includes capability and accountability. Processing data with a large 4 Vs utilizing a combination of conventional and computationally intelligent approaches is the core objective of data analysis.

2. Issues in Big Data Analytics

Both in industry & academia, analyzing of big data and field of data science has become more and more popular research subjects. Data science aims to analyse vast amounts of data and get knowledge from it. Future events can be predicted by combination of technologies and analysis. This section focuses mostly on unresolved problems. The main fields of research in huge data analysis are the (IoT) i.e., Internet of Things, cloud, bioinspired and computing of quantum.

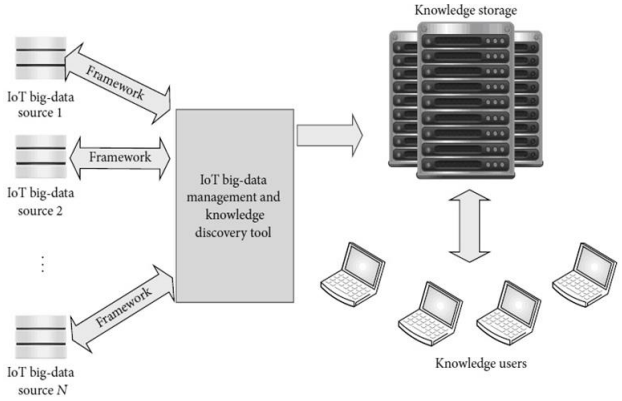

Fig. 1.

3. Big Data Processing Tools

Large volumes of data can be processed using technologies. We will mention numerous modern approaches to analysis of big data with focus on three crucial new mechanisms: Storm, Map-Reduce, and Apache -Spark. Batch, stream, and interactive analysis are the main objectives of the majority of the technologies available.

The Apache Hadoop architecture is the foundation for the majority of batch processing tools, including Mahout and Dryad. Examples of expansive streaming platforms include Strom and Splunk. Users can engage in real-time interaction with the interactive analysis process for their own analyses.

Examples of massive data platforms that support interactive analysis include Dremel and Apache Drill. We can create big data projects with the help of these tools.

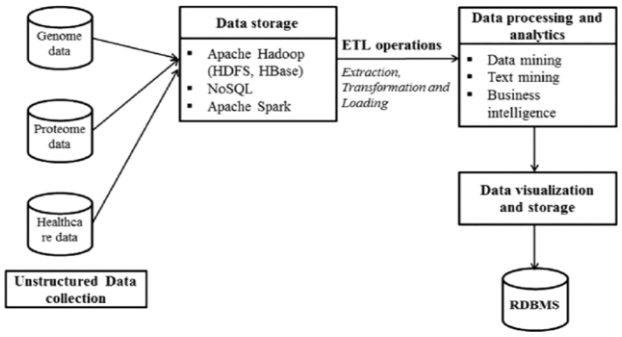

Fig. 2. Workflow of Big data project

4. Future Works

The data gathered from wide range of applications across the

*Corresponding author: divyasnaik.1727@gmail.com

globe in a number of industries may quadruplicate every 2-3 years. Without evaluating these to produce meaningful information, it is useless. Techniques that will facilitate huge data analysis must be developed in order to address this. Automated systems are the result of these ideas being put into practice more easily due to the development of powerful computers. Transforming data into knowledge is by no means an easy task for high-performance large-scale data processing, including harnessing scalability of existing and forthcoming computer architectures for data mining, turning data into knowledge is by no means an easy process.

Additionally, there could be different degrees of uncertainty in this data. Depending on the application field, big data is typically condensed to just the key characteristics needed for a given study. As a result, reducing strategies have been devised. The information gathered frequently contains incorrect or unknown values. Hence correctness of the values collected should be ensured before the data is analyzed.

The productivity, effectiveness, and adaptability of specialized data applications in computer systems may be jeopardized by these new problems, if not worse. The latter approach can cause information loss; hence it is not advised. In the business and academic communities, this poses a number of research challenges, such as successfully acquiring and accessing data. Fast processing while retaining high performance and throughput, as well as effectively storing data for later use, present additional challenges. Programming for the analysis of massive amounts of data is also a big difficulty. It is vital to specify application, requirements for data access and create abstractions in programming languages to benefit from parallel processing.

Additionally, academics are becoming more interested in machine learning theories and technology as a way to provide accurate outcomes. Big data machine learning research has mainly concentrated in the areas of processing of data, application of algorithms, and development. Many recently introduced large-scale machine learning methods need major adjustments. Although each tool has its own set of advantages and disadvantages, we think that more effective approaches to overcoming the challenges presented by huge data can be devised. Inconsistency and ambiguity, missing data, and noisy and unbalanced data must all be dealt with by the effective tools that are being developed.

5. Conclusion

In recent years, data production has increased dramatically. For a common individual, analyzing these data is difficult. In order to do this, we examine many research problems, difficulties, and analytical techniques in this study. It is clear from this study that each platform for big data has a particular emphasis. While few excel in real-time analytics, others are better suited for batch processing. Additionally, each big data platform offers particular features. Natural language processing, statistics, machine learning, data mining, data fusion and data integration are some of the several methods utilized for the analysis. We think that in the future, academics will focus more on these methods to effectively and efficiently address big data issues.

References

[1] Sanchez, D., Martin-Bautista, M.J., Blanco, I., Torre, C.: Text Knowledge Mining: An Alternative to Text Data Mining. In: IEEE International Conference on Data Mining Work-shops, pp. 664–672, 2008.

[2] Serrat O., Social Network Analysis. Knowledge Network Solutions, 28, 1–4 (2009)

[3] Shen, Z., Wei, J., Sundaresan, N., Ma, K.L.: Visual Analysis of Massive Web Session Data. In: Large Data Analysis and Visualization (LDAV), pp. 65–72, 2012.

[4] Song, Z., Kusiak, A.: Optimizing Product Configurations with a Data Mining Approach. International Journal of Production Research, 47(7), 1733–1751, 2009.