International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 4107–4113

Big Data Analytics: Applications and Challenges

Vandna Dahiya[1], Dr Sandeep Dalal[2]

[1]Research Scholar, Maharshi Dayanand University, Rohtak, Haryana.
[2]Assistant Professor, Maharshi Dayanand University, Rohtak, Haryana.
E-mail: vandanadahiya2010@gmail.com[1]
E-mail: sandeepdalal.80@gmail.com[2]

Abstract

Data analytics is the discovery or interpretation of meaningful patterns and knowledge from raw data, and converts it into valuable information, which helps in effective decision-making. With the onset of the digital era and new technologies such as IoT and cloud computing, a huge repository of data in terabytes is generated daily. Analysis of such amounts of data requires advanced analytical tools to gain knowledge for better choice making. Hence, big data analytics is the current area of research and development. The analytics in big data means joining the dots and plotting the relationships among them. This paper aims to summarize various issues in the analytics of big data.

Keywords*: Data Mining, Big Data, Distributed computing, Hadoop*

I. INTRODUCTION

Data has increased in a large scale over the last decade. For this global and explosive increase in data, the phrase 'Big Data' is used. This data is mostly unstructured and requires real-time analysis as it brings new opportunities in various fields. The basic meaning of the saying big data is that the whole lot we do today leaves a digital outline, which can be used and analyzed. It, therefore, denotes our capability to mine knowledge from the ever-increasing size of data. According to Google, "Five exabytes of data have been generated until 2003 since the dawn of civilization. Now, several exabytes of data is produced daily with the accelerating pace." Our smart-watches and smart-phones accumulate data on how we use them and what we do in our daily life. Digital music players, eBooks, online payment methods collect the data based on our activities. The food apps know our eating preferences. The cab apps and GPS know our visiting details. The browser on our phone collects all the information that we search on the Internet. All our activities and conversations through phone are recorded digitally. With the Internet of Things or IoT, we have sensors in almost all our devices. The datafication provides tremendous amounts of data with high volume, velocity, variety, and veracity. The moderntechnologies such as parallel computing and cloud-based data computing provide us the dominanceto mine this big data to gain meaningful insights and value from it. Big data provides us the prospects for an in-depth understanding of hidden values and encounters new confronts such as how to manage and process such large datasets.

ISSN: 2233-7857 IJFGCN

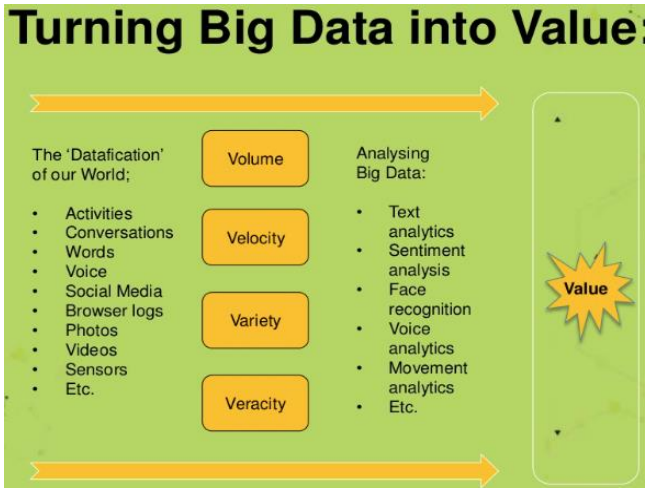

Figure 1 Big Data Analytics

II. LARGE DATA SETS - BIG DATA

Why actually the large data sets are called big data? The answer can be given from the definition of distinctive Vs, which were introduced by Gartner analyst Doug Laney in 2001.

Volume

Eight hundred hours of videos are being uploaded on the Internet in one minute. 500 million people are the daily active users of Instagram who upload photos, videos, and stories. With the arrival of IoT, terabytes of data is confined in daily by the sensors. Data is exploding in terms of volume and this size is very large to be processed by conventional computing.

Variety

Sources of data are heterogeneous, so data come from these sources can be in a variety of formats such as text, media, tables, etc.

Velocity

The continuous growth of data with a high pace is the major challenge. User behavior is captured in terms of millions of events per second.

Apart from these three, there are other aspectsalso, which define the big data. IBM in their Inforgraphic demonstrates another Vs.- Veracity and Value

Veracity

The data is captured from a variety of sources. The trustworthiness and quality of data is a matter of alarm. There would be lot of missing values, errors, noises, and biases. There is a lot of uncertainty as the data may be incomplete and inconsistent.

Value

To find some valuable information from huge data is the overall goal of big data analytics, what knowledge can be extracted, and how it can be used in decision-making and business processing.

ISSN: 2233-7857 IJFGCN

Figure 2 Vs of Big Data

III. CHARACTERISTICS OF BIG DATA: HACE THEOREM

HACE theorem depicts the key characteristics of big data. The theorem defines the big data as generated from Heterogeneous; Autonomous sources with distributed and decentralized control, and aims to investigate Complex and Evolving associations among data. In a native sense, the theorem can be understood in terms of a giant elephant and some blind men around, who are trying to size up the elephant. The aim of the men is to figure out the information from the part of the information they get. As eachmen will be limited to his own belief, different results are expected from them as shown in the figure. Exploring the elephant (big data) here will require having the best possible solution from all the illustrations, which is certainly not simple (due to diverse sources of information).HACE characteristics are briefly discussed as:

Huge and Heterogeneous Data sources

One of the characteristics of big data is the huge volume, which comes from heterogeneous data sources with diverse dimensionalities. Various sources generate data in different schemas and the sensors record them accordingly. There are different representations for the same thing. It is a major challenge to combine data from different sources.

Autonomous And Distributed Sources With Decentralized Control

Each data source, being autonomous is able to generate and collect the data without interference from other sources. There is no centralized control. It is also due to differences in demographics and government control. The sources of data are spread over the globe.

Complex Data and Evolving Relations

With the volume, complexity and the relationship among data also increase. Because of different temporal, spatial, and other factors, the features used to represent an entity are also different and complicated. The data is structured, semi-structured and unstructured. The key is to take concern of evolving changes and data relationships to find valuable patterns and knowledge from big data.

ISSN: 2233-7857 IJFGCN

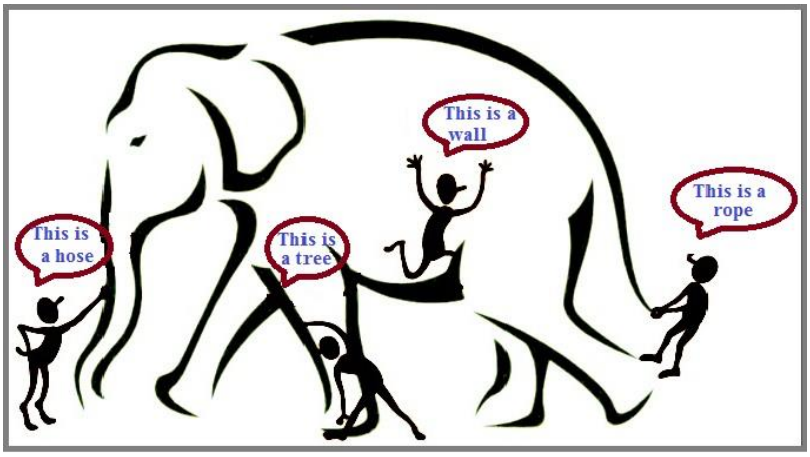

Figure 3 The Blind Men and the Giant Elephant

IV. BIG DATA ANALYTICS – APPLICATIONS

Large data sets can be analyzed to mine the patterns, trends, particularly related to human interaction. Various IT companies are investing in the field of processing and maintaining big data. Selling products and services to existing customers and making new customers by cross-selling, market basket analysis, up-selling, etc. Big data applications can be seen in various fields such as banking, agriculture, medicine, education, etc. Some of the areas are reviewed here-

Advanced Healthcare

Healthcare can be revolutionized with the potential of big data. Our wearable smart devices like watches, bands, etc. are the new personal stethoscopes. A premature sign of any abnormality can be calculated well in advance and healthcare can be informed in the era of IoT. A baby is being traced from pregnancy onwards. Based on various data sources like genetic history, medical records, and equipment, a patient is analyzed well in advance and treated accordingly. Insurance companies gain insights from our personal data and offer custom-made medical services.

Web-Analysis and E-Commerce

Based on the click history of a person, business firms can treasure out different sentiments of people. Whether there is some new product launch or a new ad campaigning, they get insights about the adores and aversions of people. Based on various click-stream analysis, link exploration, time spent on a certain page and portion of a page, they mine the thinking of the user. Target ads and applications are offered based on such mining. Match-making organization works on the same analysis. They mine the interests based on our data from social media through web mining and graph mining and then offer appropriate matches.

Market basket Analysis and Cross Marketing

Trade sectors mine the customer sales data and can arrange their stores and products accordingly. They offer various discounts on the combination of products based on utility mining of items. Cross marketing is done based on preferences of the user such as if a customer buys tickets for a comedy movie, he may be offered another series or books of the same genre on heavy discounts.

Finance Sector

The finance sector such as banks and loan firms mine the customer data and then offer their services. Based on various shopping behavior, credit-debit card analysis, and credit risk scoring, etc.

ISSN: 2233-7857 IJFGCN

risk for frauds or money laundering is calculated well in advance as predictive mining.

Media and Entertainment

Tailored services are presented to the customers based on the real-time voting system and analysis. By predicting the demands, personalized content is specified at their suitable time and according to their interest. For example, Netflix, Hotstar, Amazon prime, etc. offer the content to a user based on his/her web search history.

Telecommunication

The telecommunication sector is one of the early adopters of data mining. Customers are segmented based on customer's location, call details, usage of services, etc. and then services are proposed. Any fault in-network is also detected by mining the network statistics.

Education

Performance of students is predicted well in advance based on patterns of learning. Learning analytics is very important and targeted interventions can be provided to help them accomplish better outcomes. Feedback can be provided to students when they are stressed. Learning process can be made relaxed and personalized by designing various online courses.

V. TOOLS FOR BIG DATA ANALYTICS

With the phrase 'Big Data', the attention goes to the aspect of volume. The big data analytics require special tools and technologies to mine the data that is in exabytes. The parallel computing platform is required, which can scale as per the requirement. With the aspect of velocity, fast and intelligent architecture is required that can perform adaptive and real-time analytics. With the dimension of variety, qualitative and quantitative predictive techniques are required to integrate discoveries from unstructured data. Advanced analytics architecture (AAA) is required as the overall platform to integrate all the above aspects of big data. Some of the tools for big data analytics are discussed below.

Apache Hadoop: Hadoop is an open-source software and composed of MapReduce and HDFS-Hadoop Distributed File System. It is reliable, scalable, and supports distributed computing ranging from a single server to thousands of servers. It divides the task into small jobs and assigns them to different nodes.

Apache Mahout: It is a project of Apache Software Foundation and used for distributed computing and scaling for machine learning algorithms, mainly that use liner algebra and common Mathematics operations.

Apache spark: This software is also from Apache Software Foundation. The difference in Hadoop and Spark lies in several facts. Here, mappers are reducers are independent of each other. Processing is done in-memory without repeating read/write operations, which makes Spark faster and efficient than Hadoop.

Storm: Storm is a free and open-source distributed computation system that is used for processing of streams. Streams are partitioned for various stages of computation for real-time processing. It can be used with any language.

Apache Drill: It is an SQL query engine and used mainly for the processing of semi-structured data. It achieves low-latency for queries of SQL over large datasets.

ISSN: 2233-7857 IJFGCN

VI. CHALLENGES WITH BIG DATA ANALYTICS

Analytics is a multidimensional field, which uses machine learning, mathematics, statistics, and predictive modeling. Most of the mining and analytical algorithms were developed before the era of big data and hence why they don't consider the features of big data such as large size, velocity, etc. Reliable and scalable data mining algorithms are required to extract knowledge from such data sets. There are various other requirements that need to fulfill by data mining algorithms. These are briefly discussed in this section.

Meeting the needs for Volume and Speed

Organizations not only need to analyze the huge size of data but also need to extract the information as fast as possible taking into consideration the competitive environment. With an increase in granularity, the challenge also increases. Possible solutions are either using hardware with multiple processors and expanded memory or grid computing like architecture where multiple computers are used to solve a particular problem.

Addressing the quality

The information mined should be accurate and in stipulated time so that it can be useful for the end-user. Data quality should be assured to make the whole process of analysis useful. To address the issue of quality, there should be some governance that can ensure that data is consistent and accurate. Suitable pre-processing is required to maintain the cleanliness of the data.

Visualization of Results

The output of big data mining plays a significant role. There should be additional metrics such as data loading time, mapping-reducing time, time for queries, etc. apart from standard metrics like memory usage, execution time, and accuracy. To analyze the results fully by the user, it is important that they should be presented in a format that a user can understand. For example, if there is a need to compare users of Instagram and we try to plot the billion of dots on graph, it would be impossible for the user to gain any insight. For that matter, the clustering of data with a high-level view can be used. Another factor can be outliers. Outliers represent around 5% of the total output. To plot the outliers on the graphs or other visualization tools might be difficult when working with big data. Also, the user might find it difficult to understand these points on graphs. Separate graphs for outliers should be used for better visualization.

Privacy and Security

Although the sources of data are autonomous, there is a concern of privacy and security whenever the personal information of a user is involved. Locations of a user, preferences, daily activities, etc. are stored in the systems. There should be some security interface to prevent such information from vulnerable attacks.

CONCLUSION

Big data analytics is a budding area of research. In this paper, the current status is presented with respect to characteristics, various tools, and challenges of big data. Big data can deal the marvelous insights into the companies. But with huge amounts of data daily bucketing in an organization, the conventional computing infrastructure is not suitable and up to the mark. An integrated-advanced analytical architecture is required for real-time applications for quick insights. It is vital to recall that the value from big data does not come from raw data but with the insights that surface from processing and analysis of it.

ISSN: 2233-7857 IJFGCN

REFERENCES

1. X. Wu, X. Zhu, G. Q. Wu, and W. Ding, **"Data mining with Big Data,"** Knowledge and Data Engineering, IEEE Transactions on, vol. 26, no 1, p. 97-107, 2014.
2. A. Gandomi and M. Haider, **"Beyond the Hype: Big data Concepts, Methods and Analytics",** International Journal of Information Management, vol. 35, no. 2, pp. 137–144, 2015.
3. D. H. Shin and M. J. Choi, **"Ecological Views of Big Data: Perspective and Issues",** Telematics and Informatics, vol. 32, no. 2, pp. 311–320, 2015.
4. C. Magdalena, R. Martínez-Espana, B. Ayuso, J. Antonio Yanez, A. Munoz, **"Analysis of Student Behavior in Learning Management Systems through a Big Data Framework",** Future Generation Computer Systems, vol. 90, pp. 262-272, 2019
5. Sandeep Dalal, Vandna Dahiya, **"Big Data Mining: Current Status and Future Prospects,"** International Journal of Advanced Science and Technology. 29, 3 (Mar. 2020), 4659- 4670, 2020
6. Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, **"Big Data Analytics,"** Journal of Big Data, Springer, Vol. 2, No. 21, October 2015. Pp 13-52
7. Deepak S. Tamhane, Sultana N. Sayyad, **"Big Data Analysis Using Hace Theorem",** International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 1, January 2015, pp 18-23
8. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, **"Data Mining with Big Data",** IEEE Transactions On Knowledge And Data Engineering, Volume 13, (4), 2013, pp 1-24
9. J. Manyika et al., **"Big Data: The Next Frontier for Innovation, Competition and Productivity,"** New York, NY, USA: McKinsey Global Institute, 2011.
10. A. Intezari and S. Gressel, **''Information and Reformation in KM systems: Big Data and Strategic Decision-making,''** J. Knowledge Management, vol. 21, no. 1, pp. 71–91, 2017.
11. D. P. Acharjya and A. P. Kauser**, ''A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools,''** Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 2, pp. 511–518, 2016.
12. Dalal Sandeep, Dahiya Vandna, **"Review of High Utility Itemset Mining Algorithms for Big Data",** In: Journal of Advanced Research in Dynamical and Control Systems- JARDCS, 10(4), pp: 274-283, 2018
13. Guangming Guo, Lei Zhang, Qi Liu, Enhong Chen, Feida Zhu, Chu Guan, **"High Utility Episode Mining Made Practical and Fast,"** Advanced Data Mining and Applications,pp 71-84, 2014, Springer, 2016
14. G. Manogaran, C. Thota, D. Lopez, V. Vijayakumar, K. M. Abbas, and R. Sundarsekar, **''Big Data Knowledge System in Healthcare,''** in Internet of Things and Big Data Technologies for Next Generation Healthcare, C. Bhatt, N. Dey, and A. S. Ashour, Eds. Springer, 2017, pp. 133–157.
15. http://hadoop.apache.org
16. http://www.philippe-fournier-viger.com/spmf

ISSN: 2233-7857 IJFGCN