International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-8, Issue-3, May 2020
https://doi.org/10.21276/ijircst.2020.8.3.29
www.ijircst.org

Big Data Analysis: Concepts, Challenges And Opportunities

Shweta Sinha

ABSTRACT- Big Data has the potential to bring valued insights for advanced decision making. It has become a trending practice to adopt Big Data into their process. It has attracted both the academicians as well as the industry people. A voluminous amount of diverse data is generated every day, and Big Data has emerged as an approach to process them. Data can be analysed in a different manner depending upon the requirement. The paper identifies the challenges and issues associated with Big Data and points out the research and industry opportunities in this field.

KEYWORDS- Big Data Analytics, Big Data Challenges, Big Data Lifecycle, Big Data Opportunities.

I. INTRODUCTION

Big Data is a domain dedicated to the storage, analysis and processing of enormous data size that is collected from disparate sources. When the traditional ways of storage analysis and processing fail to handle the data size, Big Data comes into the picture. It can be said that Big Data is needed to fulfil some specific requirements such as handling of unstructured data, combining unrelated data of heterogeneous nature as well as extracting hidden information. The management and analysis of large datasets have been a long-standing problem—from labour-intensive approaches to the actuarial science behind the calculations of insurance premiums. As the data size growth is tremendous, the big data techniques leverage computational resources and strategies to execute analytics. This shift is a welcome step in this direction.

The analysis of Big Data datasets is an interdisciplinary endeavour that blends mathematics, statistics, computer science and subject matter expertise. The potential of Big Data is evident and has been included in the Top 10 Critical Tech Trends for the Next Five Years [1]. It is as vital as nanotechnology and quantum computing in the present era. In essence, Big Data is the artefact of individuals as well as collective intelligence that are generated and shared mainly through the technological environment. The volume, variety, and dynamism of such data demand a new type of analytics, along with a different storage method. Such total Big Data need to be correctly analysed, and information embedded should be extracted.

Manuscript received May 22, 2020

Dr Shweta Sinha, Associate Professor in the Department of Computer Science and Engineering, Amity University, Gurugram Haryana, India (email: shwetas.kiit@gmail.com)

The contribution of this paper is to provide an insight into the analytics of Big Data. Section 2 covers the characteristics of Big Data, followed by Methodology in section 3. Section 4 presents the assessment of tools and section 5 discusses the application domain of Big Data and is followed by section 6 in conclusion.

II. BIG DATA ANALYTICS

The Big Data analytics lifecycle generally involves identifying, procuring, preparing and analysing large amounts of raw, unstructured data to extract meaningful information that can serve as an input for identifying patterns, enriching existing enterprise data and performing large-scale searches. There are four general categories of analytics that are distinguished by the results they produce; descriptive analytics, diagnostic analytics, predictive analytics and prescriptive analytics. Different analytics types leverage different techniques and analysis algorithms. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualising. Today, enterprises are exploring large volumes of highly detailed data so as to discover facts they didn't know before [2]. With the evolution of technology and the increasing multitudes of data flowing in and out of organisations daily, there has become a need for faster and more efficient ways of analysing such data. Having piles of data on hand is no longer enough to make profitable decisions at the right time. Different kinds of organisations use data analytics tools and techniques in different ways. Some data sources, such as sensor captured data, can produce staggering amounts of raw data. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude [3]. Regardless of where Big Data is being generated from and shared too, with the reality of its existence comes the challenge of analysing it in a way that brings big value. Owing to its importance and the value residing inside, it is regarded as today's Digital Oil [4]. Appropriate data processing and management could retrieve new knowledge and help in responding to emerging opportunities and challenges in a timely manner [5]. Established data processing technologies such as database and data warehouses are becoming inadequate due seeing to the amount of data the world is currently generating. This humongous data needs to be analysed in a repetitive fashion, as well as in a time-sensitive manner [6]. With the availability of advanced Big Data analysing technologies namely, NoSQL Databases, BigQuery, MapReduce, Hadoop, perceptions and understandings can be better achieved to enable in improving the business

policies and the decision-making process in many critical sectors such as healthcare, economics, energy futures, and predicting natural catastrophe, to name a few [7]. Data Analytics can help strengthen and squeeze the focuson delivering high-quality services by driving the costs down.

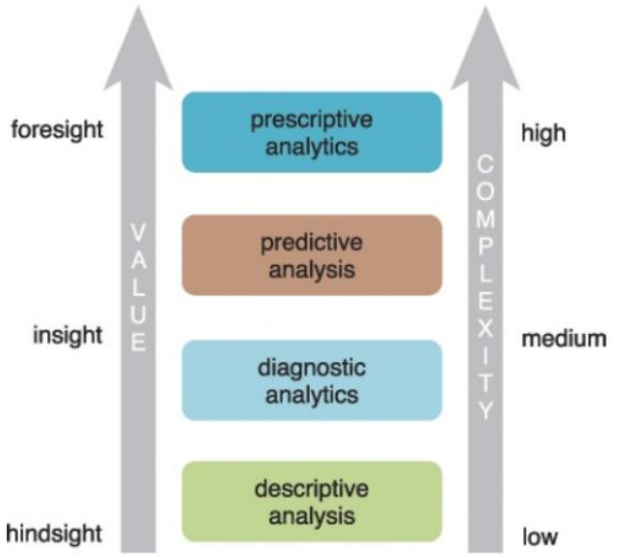

Fig 1: Value and complexity relation with analysis type [8]

- Descriptive Analytics: It is the simplest form of analytics applied to the collection of data. This involves the summarization and description of knowledge patterns using simple statistical methods, such as mean, median, mode, standard deviation, variance, and frequency measurement of specific events in Big Data streams [9]. This type of analytics is often done for a large volume of historical data to analyse hidden patterns and present a report on that.
- Diagnostic Analytics: Diagnostic analytics determines the cause of a phenomenon that has occurred in the past. This is done using questions that focus on the reason behind the event. The goal is to determine what information is related to the phenomenon in order to enable answering questions that seek to determine why something has occurred.
- Predictive Analytics: These analytics is concerned with forecasting and modelling based upon supervised/unsupervised learning from the history data. These analysis techniques are primarily based on statistical methods that are used to discover the relationship between data and uncover hidden patterns within the data.
- Prescriptive Analytics: This type of analytics is performed to determine the cause-effect relationship between analytic results and business process optimisation policies. Thus, for prescriptive analytics, organisations optimise their business process models based on the feedback provided by predictive analytic models [10]. The result of the analytics is dependent upon the characteristics of data gathered. If the limited dimension of data is captured howsoever secure the predictive analytic model be the result will be constrained to finite size only.

A. Big Data Characteristics

Any data can be considered as Big Data if it possesses some characteristics that require to be present in the architecture and solution data of analytics. In early 2001, owing to the nature of data that needed some special attention for analysis, the four V's were defined to characterise them. These were mainly focused on the characteristics of e-commerce data. Sooner it was realised that the collected data, not all fall into the same category and the data could be structured, semi-structured or unstructured. To cover the characteristics of all these a new V was added and then the 5 V's were defined as:

- Volume: Volume refers to the large amount of data being generated and processed. This high-volume data impose distinct data storage and processing demands and also some additional preparation and management process.
- Velocity: In the Big Data environment, data accumulates at a very high speed, and in a very limited time, it can reach a huge volume. Velocity refers to the time needed for data processing once it enters the system. Coping with fast inflow requires special solutions to handle it.
- Variety: Variety in Big Data refers to the multiple formats in which data arrives from different sources. These data can be either any tabular data or textual data and may also include audio, video, XML files and many more. Dealing with these requires some mechanism for integration, transformation and processing.
- Veracity: It refers to the quality or the stability of data. Data may arrive from wrong sources and can be noisy data with false information. Before including them into the process, it has to be checked for its correctness and mechanism has to be applied for removal of noise or incomplete data.
- Value: Value is the measure of the usefulness of data for any enterprise. It is somehow related to the veracity and the timeliness of processing. The higher the data correctness, the higher will be its usefulness and in turn, value. Also, the in-time process of data has its value, whereas delayed processing discards the data. It is essential to have a processing characteristic that handles these issues.

B. Big Data Life Cycle

i. Data collection – The first phase starts with data collection. Identification of data sources is the most crucial aspect. More data sources will lead to more information relating to correlation and stability.

ii. Data storage- The captured data can be in structured or unstructured data, and for further processing, one need to store them in a database like NoSQL or other that supports such storage and processing.

iii. Noise Elimination- Noise elimination deals with the cleaning of data. This requires removal of duplicate, null and irrelevant data from the gathered information. It is always advisable to keep the original copy intact.

iv. Data classification and extraction- The next element of the cycle deals with the extraction of relevant data from the cleaned source. This also helps in the reduction of the volume. Several tools can be used for this purpose.
v. Data validation and aggregation- During this stage business rules are identified, and the data is validated for its relevance with respect to the business needs. Aggregation is used to combine multiple data values into fewer numbers based on common fields and simplifies further data processing.
vi. Data analysis and processing- This stage is responsible for carrying out actual data analysis to find out the hidden information from the data collected. The data is processed in different ways depending upon the type of analytics needed.
vii. Data visualisation and evaluation- The last stage is an important stage as it represents the overall findings for the user. The representation should be based upon the requirement and should be easy to understand. Deciding on the visualisation tool is important as it is directly related to the user's need.

III. TOOLS AND METHODS

With a high volume of data moving through the business process system and also with advancements in technology, there is a need for faster and efficient ways to analyse and process the data. The timely fulfilment of the demands in an easy to visualise the result is what exactly is needed. Traditional tools and methods were not sufficient to process under these constraints. Therefore, there arises a need for new tools and techniques specialised for big data analytics, as well as the required architectures for storing and managing such data. Developers came up with several tools for the processing of data. Table 1 presents the tools for different phases of Big Data life cycle.

Table 1: Tools for different stages of Big Data life cycle

Life Cycle Stage	Tools	Characteristics
Data Collection	Opinion Crawl	Web-based
	OpenText	Window-based application
Data Storage	Apache HBase	No Secondary index
	MongoDB	Secondary index
	Oracle NoSQL Database	No Secondary index
Data Filtering	OctoParse	Structured spreadsheets
	ParseHub	Excel, CSV, Google sheet
	Content Grabber	Structured data (XMLCSV)
Data Cleaning	DataCleaner	Record and field processing
	MapReduce	Parallel data processing
	OpenRefine	Batch processing
Data Analysis	Hive	Streamed data
	Apache Spark	Mini/ micro-batches of data
	MapReduce	Parallel Processing
	Flink	Batch and stream processing
Data Visualization	DataWrapper	Bar chart, line chart, map, graphs
	Tableau	Maps, Bar charts, Scatter plots
	CartoDB	Maps
	Gephi	Graphs and networks
	Google Fusion tables	pie charts, bar charts, line plots, scatter plots, timelines

IV. BIG DATA CHALLENGES

The large data set and ability to handle complex data sets in terms of time and complexity has always been the most challenging task for traditional systems. With time several tools and methods came into existence to handle these issues, but some of the challenges still persist. These challenges can be due to data itself or can be due to processing or managing aspects of massive datasets.

A. Data Challenges

These challenges are the ones related to the characteristics of the data itself. That is the five V's that define the information also possess one or more problems in this area.

- Volume: a large volume of datasets consisting of terabytes to zettabytes of data is a challenge to manage and process. Even if tools and methods exist for handling Big Data, the growth in volume is becoming unmanageable.
- Variety: Data is collected from diverse sources and are of diverse nature. Using these heterogeneous data for processing and integration possess a challenge.
- Veracity: Challenges arise due to the existence of biases, fabrication of data and noise in the data. Accuracy of data is influenced by these. Removing them from a huge volume of data is a big challenge.
- Velocity: Challenge arise due to high-influx of data, and that too non-homogenous data. Processing these data sets in a speed matching to the velocity of its generation is a difficult task.
- Variability: The data on the internet are constantly and repeatedly changing. Using them to conclude or predict something is a challenge as it may lead to inconsistency of the report.
- Value: There are always lots of information within the data

that has been left unused on the internet. Multiple dimensions of data need to be analysed to extract value from them. Regardless of the number of dimensions used and availability of several tools, the challenge for extracting value from the data persists.

B. Process Challenges

These are the challenges encountered during processing and analysing the data and extend to the interpretation and presentation of information. These challenges are associated with the lifecycle process of Big Data.

- Data acquisition: Acquiring data from diverse sources and storing them to extract some value out of it is a challenge. Smart filters are needed to obtain data that are useful and discard unnecessary data.
- Data cleaning and mining: The challenge at this stage is to extract and clean data from the collected pool of large-scale unstructured data. The vibrant, diverse, interrelated features of data make the mining and cleaning task difficult.
- Data aggregation and integration: Aggregating and integrating a large pool of data captured from diverse sources need a lot of processing. Handling a huge amount is a challenge. Data from tweets, Facebook comments have diverse meaning and binding them together is difficult.
- Data analysis and modelling: Once the data has been extracted and cleaned the analysis is needed to extract value out of it. Traditional methods of analysis can not handle these faster-growing, unstructured unrelated data. Tools and resources are needed to handle them efficiently.
- Data interpretation and visualisation: The unstructured data has a multiplicity of interpretations. There is a huge shortage of manpower resources to handle these.

C. Management Challenges

These are the challenges encountered while accessing, managing and governing the data.

- Privacy: To preserve privacy in this digital age is a huge challenge. Huge investments have been made in this field, but till date, it remains the challenge. This is also pulling back the business world to leverage the advantages of Big Data for industry growth.
- Security: Security is again a significant challenge associated with Big Data. Due to this, the phenomena of Big Data have not yet been accepted globally. Most of the security concerns with the Big Data are the ones that are also associated with the traditional databases. Distributed nature of data possesses additional issues relating to network log.
- Data governance: It is an approach to warranty data quality, leveraging information and support insight into the business. The challenge is to decide what data to store, process and analyse for providing such services timely.
- Data and Information Sharing: The sharing of data and information needs to be balanced. It should be controlled to maximise its effect because this will facilitate organisations in establishing close connections and harmonisation among themselves.
- Data Ownership: Data ownership is a big issue, as is the data itself. For data on Twitter, Facebook or other social media, this type of challenge is more. Who owns the data actually stored on these platforms; the user who uploaded it or the platform/organisation that hosted it? Both the views exist in this regard and creates a challenge for coming up with a decision.

V. OPPORTUNITIES IN BIG DATA ANALYTICS

With the growth of Big Data, it is evident that lots of potential reside in this field. The opportunities that this domain provide to the analytics are:

- Real-time processing of data
- Central management of data
- Managing disparate data sources
- Leverage, Big Data stores, to obtain multiple insights
- Information for everyone and everywhere

Apart from these outlined as the opportunities, there are many more and uncountable opportunities in Big Data analytics. Its promise to handle complex datasets and produce in the simplest fashion is being fulfilled, and the efforts are ongoing.

VI. CONCLUSION

This research examines the innovative topic of big data, which has gained lots of interest due to its perceived unprecedented opportunities and benefits. In the present era, the voluminous amount of high-velocity data is generated daily. These data, when properly used, brings insight into the business processes. The analysis, challenges and opportunities of this field are presented in this paper. The paper covers the life cycle of Big data processing and presents tools for each of the stage. The aim of the review presented here is to highlight the research prospects and to motivate the readers to work in this challenging field.

REFERENCES

[1] Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V., 2017. Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, pp.263-286.

[2] Elgendy, N. and Elragal, A., 2014, July. Big data analytics: a literature review paper. In Industrial Conference on Data Mining (pp. 214-227). Springer, Cham.

[3] Labrinidis, A. and Jagadish, H.V., 2012. Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12), pp.2032-2033.

[4] Yi, X., Liu, F., Liu, J. and Jin, H., 2014. Building a network highway for big data: architecture and challenges. Ieee Network, 28(4), pp.5-13.

[5] Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S. and Zhou, X., 2013. Big data challenge: a data management perspective. Frontiers of Computer Science, 7(2), pp.157-164.

[6] Labrinidis, A. and Jagadish, H.V., 2012. Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 5(12), pp.2032-2033

[7] Sureshrao, G.S. and Ambulgekar, H.P., 2014, August. MapReduce-Based warehouse systems: a survey. In 2014 International Conference on Advances in Engineering & Technology Research (ICAETR-2014) (pp. 1-8). IEEE..

[8] Erl, T., Khattak, W. and Buhler, P., 2016. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press.

[9] ur Rehman, M.H., Chang, V., Batool, A. and Wah, T.Y., 2016. Big data reduction framework for value creation in sustainable enterprises. International Journal of Information Management, 36(6), pp.917-928..

[10] Bihani, P. and Patil, S.T., 2014. A comparative study of data analysis techniques. International journal of emerging trends & technology in computer science, 3(2), pp.95-101.

ABOUT THE AUTHORS

Dr Shweta Sinha pursued MTech from GGSIP University Delhi and her PhD from Birla Institute of Technology, Mesra, Ranchi, India. She is currently working as Associate Professor in the Department of Computer Science and Engineering, Amity University, Haryana, India. She has 15 years of teaching experience and 7 years of research experience. She has published 19 research papers in reputed journals and conferences and has four book chapters to her credit. Her main research area is speech and language processing.

