ISSN:0975-9646

A Survey on Big Data Analytics: challenges and opportunities

Bindu M G,
Department of computer Applications, Sree Keralavarma college
Thrissur, India

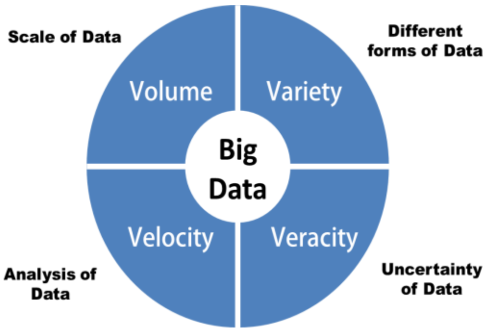
Abstract – **Digitalization has been producing enormous amount of data from various sources. These data sets are collectively called as Big Data. Big data is not only big by volume but it is high in velocity and variety. This makes traditional methods of data processing unable to handle Big Data. But for proper Decision making we need to extract useful information from the stored data. Otherwise the entire struggle to store the huge data will be in vain. There comes the relevance of Big Data Analytics. This paper provides a view to different analytical methods and tools to handle Big Data. It also explores different challenges in the area of Big data analytics. In addition, this paper gives glimpses of new research problems for the budding research scholars.**

Keywords—**Big data, review, benefits of big data analytics**

I. Introduction

The world is going through an age of "Big Data". Enormous amount of data is being produced in each second. The amount of data created annually is predicted to reach 80 zetta bytes by 2025. Storage of the data is promoted since organizations can extract useful information from the historic data and take appropriate decisions. All data may not be useful in decision making. The task is to scan through the "Big" data which is high in volume and is being produced with high velocity, to find useful information. Traditional data mining methods are found ineffective in handling huge amount of data[1]. So there is a need for new tools and techniques to handle the big data. Big data analytics is the collection of advanced analytic techniques applied on Big data for proper extraction of useful data.

Big data can not only be characterised by its volume. Size is just one of the dimensions of Big Data. It is described by 4 v's. The following figure gives the four v's[2]



Obviously, the "Big data" will be big in volume. Big data involves data produced by different devices and applications such as Social sites, search engines, Medical history, Online shopping data, Stock exchange data etc. Out of the 7 billion world population, 6 billion are using cellular phones which produces huge amount of data. It is estimated that 2.3 trillion gigabytes of data is being produced in each day[3]. In each passing minute the volume of data is cumulatively increasing. By 2020, 43 trillion gigabytes of data will be produced which is 300 times the amount of data produced in 2005[3]. The speed at which data is being produced is known as the velocity. Beyond 2020, many sources predict an exponential growth of data. Decisions are to be taken dynamically as the data is evolving with a higher speed. Only such decision will be beneficial to the organizations.

The data produced can be structured, semi structured such as XML or unstructured such as text, human language since the data is generated from diverse resources. This brings variety in the big data. Veracity refers to the trustworthiness of the data. Since the data is available in big volumes there can be junks too. So Decision taken on behalf of information analysis should handle the reliability factor too. According to IBM's Big data and analytics hub, one in three managers are reluctant to trust the information to take decisions. Poor data quality can cost millions for a business organisation. In addition to the four v's , more components are added to explain big data, viz.. value, venue, vagueness, validity, vocabulary and variability.

This paper analyses the studies that have been carried out in the area of Big data and briefs them. This paper aims to help the new comers of the Big data research field to get introduced with the terminologies, developments, tools and techniques. This paper has been arranges as three major sections. Out of which the first section deals with the challenges in the area of Big data analytics, the second handles the researches happened so far and the issues being raised and the third deals with the tools and techniques used to process the big data. The paper concludes by throwing some light to some upcoming research areas in big data.

II. Challenges in Big data analytics

Nowadays Big data finds application in almost all areas of human life and so does in different applications. Big data has given its benefits in the fields of health care, social computing, education, administration, bioinformatics etc. Upcoming researchers can always use these opportunities. With opportunities there always come challenges. Various researches had been carried out to handle different challenges in many areas of applications. Broadly the challenges fall in the following major categories.

- Practical issues in big data storage and analysis
- Difficulty in data analysis and computational complexities
- Synchronization Across the Data Sources
- Information security and privacy.
- Practical issues in big data storage

Since there is huge amount of data to store, traditional storage systems are insufficient in storing it[4]. Also existing data storage systems are not able to give adequate performance in the Input /Output processing. Hence new storage system has to be designed which gives proper performance while handling big data.

- Difficulty in data analysis and computational complexities.

Big data promoted the usage of innovative data management frameworks which incorporated both operational and analytical processing. Platforms like NoSQL have been developed so as to support the performance demands of Big Data. There exist a wide variety of NoSQL tools. Still there exists uncertainty regarding which tool to use in which scenario. Improper selection of the tool can cause economical downfall for the customers. So the technology risk is one of the challenges in the field of Big Data[5].

Analysis of huge amount of data will always bring computational complexities. Inconsistencies and uncertainties have to be dealt with using complex computations involving deep learning techniques. Much research has been carried out to incorporate machine learning with minimum memory requirements [6].

- Synchronization Across the Data Sources

Big data has a characteristic of variety. Heterogeneous data is difficult to manage. Since the data is being collected from various resources in various format, it is not easy to ensure synchrony in Big data. Since the speed of data production and updation is immense, the lack of governance may lead to serious inconsistencies in the information. An asynchrony in big data can cause disastrous results [5]

- Information security assurance.

Different organisations keep big data for their analysis and all of them will be having different policies to safeguard the privacy of data. Since there is no proper governance for enhancing security, it has become one of the challenges in big data. Though techniques like authentication, authorisation and encryption have been implemented for general security, big data needs multi-layered security. Researches are still going on to enhance security at different levels [7] .

III. Applications of Big data

- Big data in healthcare

Nowadays millions of people are using smart phones and wearable devices for healthy life styles. With the help of big data a person's data is not only processed in isolation, but it is compared and analysed with thousands of others. Thus helps to highlight specific threats by identifying patterns [8] Prediction models based on big data can measure the variables in the body and predict the chance of occurrence of a particular disease, which leads to prevention of a disease before it spreads. Big data has thus transformed the area of health care.

- Smart farming

Big data is being used to give predictive information in farming operations so users can always make use of it to take decisions and business redesign. Rapid development in the field of cloud computing and Internet Of things are giving a new face to the farming, smart farming. Advanced machines with built in intelligence can record real time data. Big data analytic tools can be effectively used in processing this data and extracting useful information [[10].

- Fraud management

Fraud detection works by identifying certain patterns. Quick fraud detection is essential to minimize losses. The faster a bank detects fraud, the faster it can restrict account activity. The customers too will be sensitive about their financial information. Big data analytics helps in doing fraud detection fast. In addition to the traditional database data BDA makes added use of social profiles for fraud detection.

- Education

Big Data provides information to the area of Learning Analytics that allows academic institutions to better understand the learners' needs and proactively address them. Big Data and Analytics can be applied to various settings within higher education such as administrative and instructional applications, recruitment, admission processing, financial planning, donor tracking, and student performance monitoring [9]

- Retail

Nowadays, customers have 24 hours access to abundant product information. This caused a revolution in the field of Retail. With digital technology being popular, people can buy anything from anywhere at any time. Retailers are depending much on data analysis to predict customer behaviour. Big Data analytics is now being applied at every step of the retail process - right from predicting the popular products to identifying the customers who are likely to be interested in these products and what to sell them next. It works by generating recommendations to the customer depending on his purchase history. It also predicts the market trends by listening to the social media and helps in optimising the prices.

- Communication, Media and entertainment

Organizations in this industry largely analyse customer data and behavioural data to create customer profiles, which can be used to create content for targeted audience, recommend content etc. Different applications make use of sentiment analysis to rate content.

IV. Tools and techniques

With the evolution of technology huge amount of data is flowing through different devices. To handle the data properly and make it useful new tools and methods have to be developed. There comes the relevance of Big Data Analytics and Decisions (B-DAD) framework which incorporates different tools and methods for data storage and management, processing and analytical tools, visualisation tools for used in different phases of decision making. The main areas where the tools and methods are applied are storage and management, data and analytics processing, big data analytics.

Storage and management

The main concerns while dealing with big data are where and how to store it. The traditional approaches of data storage and management include relational databases, data warehouses etc.. which use ETL(extract, transform ,load)mechanism to manage data. But big data environment demands a Magnetic, Agile and Deep(MAD) mechanism to handle it. The mechanism should be magnetic so that it should attract data from different resources irrespective of its quality. Since the data available may not be purely structured the storage mechanism should have provision to handle the semi-structured and unstructured data. It should adapt with the rapidly growing data so should be Agile. Since the big data analysis involves complex statistical algorithms which makes the analyst go up and down the data sets, the data repository should be deep enough.

To satisfy different demands to deal with Big data, several solutions ranging from Massive Parallel Processing (MPP) databases to non-relational or in-memory databases have been used for big data. Non-relational databases such as NoSQL were developed for handling non-relational, unstructured data. NoSQL aims at high scaling, simplified application development and deployment. In-memory databases handle data in server memory thus reducing the overhead of disk input output processing. It provides real time data access to database.

Hadoop, an open source framework written in JAVA is used to process high volumes of data in any structure. It allows distributed storage and distributed processing for huge data. It has two components-

1. Hadoop distributed file system for big data storage
2. MapReduce for big data processing

In HDFS, a single file in split into blocks and distributed across cluster nodes. Data is saved using a replication mechanism to deal with node failures. HDFS has two types of nodes: Data nodes (to store data)and name nodes(act as a regulator between client and data node).

Data and analytics processing

Data analytic processing faces different challenges such as fast data loading, fast data querying, efficient utilization of storage processes etc. MapReduce is a parallel programming model which breaks a task down into different stages and executes the stages in parallel to save the time for entire computation process. First phase of MapReduce is to map different input values to a set of different key-value pairs . The map function partitions large task into smaller tasks and assign them to key/value pairs. This is the input for the reduce function. Reduce combines the keys which carries the same values and forms the output. Hadoop stores data in distributed files and parallel runs MapReduce for computations.

Big data analytics

Once the big data is stored and processed, it will be easy for the decision makers to work on it to extract useful information. Data analytics is applying algorithms on big data sets to analyse and extract useful information from the data. Analytics reveals the unknown or unrecognised patterns from big data sets which disclose important relationships between stored variables. Thus big data analytics attracts the attention of researchers. In addition to the traditional analytic methods like clustering, decision trees and regression, big data has brought new methods such as social network analysis, opinion mining or sentiment analysis, advanced data visualisation and visual discovery.

Social media analysis is more than analysing who shares what on the network. It analyses the reaction and conversation between people online and extract useful patterns from them. This is done using text mining or sentiment analysis. Social network analysis differs from social media analysis and focuses on relationships between social entities. It also facilitates the flow of information between interacting entities.

Sentiment analysis analyses and understands the emotion pattern from a subjective text available on FaceBook, Twitter or any blogs. Sentiment analysis uses natural language processing and text mining to understand opinion and emotion of people towards some topics.

Advanced data visualisation technique is a powerful tool to discover knowledge from data. It can be used to analyse data at both introductory and detailed level. It is a more effective presentation, analysis and decision making tool. It can accommodate a large set of data from different sources which makes it a perfect fit for big data analytics.

V. Recent Research Trends and future scope in Big data analytics

Big data has become the focus of recent researches. Organisations are paying attention to store and process the Big data to extract useful information for them. Big data finds applications in wide variety of fields. Effective integration of different technologies and their analysis can produce more effective results. Research issues in big data touches the fields of health care, cloud computing based analytics, machine learning, Internet Of Things for big data analytics etc. Some of them are discussed in detail.

Bio inspired computing is an emerging research area in big data analytics, which uses nature-inspired computing in solving real world problems. The techniques used in bio molecules like DNA and proteins to store, retrieve and process data are copied for complex calculations. Bio inspired computing is used in intelligent data analysis. It has got better interactions, minimum data loss and better ways to handle ambiguities.

Virtual computing has made most of the complexities in computations easily possible. Cloud computing manages massive data demand by giving access to computing resources through virtualization techniques. The benefits of Cloud computing include offering resources when there is a demand and pay only for the resources which is needed to develop the product. Simultaneously, it improves availability and cost reduction [7] The challenges of big data analytics are mostly based on its velocity and variety. Traditional tool are insufficient to manage the huge data. Cloud computing offers a model for solving all these challenges. Cloud computing also serves space for storing the big data.

Machine learning tools are gaining attention among research scholars and are being widely accepted by them. Deep learning concepts like fuzzy set, rough set, their generalisations and hybrid models can be associated with Big data to build more efficient predictive models. Though researches have been carried out in this realm, much more have to be done. In-memory analytics is another stream which offers a bundle of research opportunities. It is simply remembering the past and looking forward at the future. These are just glimpses of trends and future research scopes. Much more has happened and are yet to happen.

VI. Conclusion

In recent years the amount of data has exploded to make it really tough to handle. Proper storage and processing of this data yields better assistance for decision making. Thus handling this enormous amount of data has to be studied in detail and techniques have to be developed. Big data finds applications in several fields which are close to our day to day life. The growth of data is expected to be rigorous in the years to come. This paper goes through the fundamentals, challenges, tools and research trends in big data. It concludes by investing the hope in the upcoming studies which will improve Big data analytics and make it more popular.

References

1. A SURVEY OF BIG DATA ANALYTICS Nirali Honest1 and Atul Patel2 - International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March 2016
2. http://vusumuzi.dbsdataprojects.com/2017/03/04/big-data-and-analytics/
3. IBM's Big data and analytics hub. Available at http://www.ibmbigdatahub.com/
4. Duggal, Reena & Balvinder, Shukla & Khatri, Sunil Kumar. (2016). Opportunities and Challenges of Using Big Data Analytics in Indian Healthcare System. Indian Journal of Public Health Research & Development. 7. 238. 10.5958/0976-5506.2016.00226.6.
5. https://www.progress.com/docs/default-source/default-document-library/Progress/Documents/Papers/Addressing-Five-Emerging-Challenges-of-Big-Data.pdf
6. O. Y. Al-Jarrah, P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and K. Taha, Efficient machine learning for big data: A review, Big Data Research, 2(3) (2015), pp.87-93.
7. A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools D. P. Acharjya School of Computing Science and Engineering VIT University Vellore, India 632014, Kauser Ahmed P School of Computing Science and Engineering VIT University Vellore, India 632014
8. Applications of big Data: Current Status and Future Scope 1 Sabia, 2 Sheetal Kalra Department of Computer Science & Engineering Guru Nanak Dev University Regional Campus, Jalandhar, India
9. Kalota, Faisal. (2015). Applications of Big Data in Education. International Journal of Social, Education, Economics and Management Engineering. 9. Kalota, F. (2015). 'Applications of Big Data in Education'. World Academy of Science, Engineering and Technology, International Science Index 101, International Journal of Social, Education, Economics and Management Engineering, 9(5), 1501 - 1506..
10. Big Data in Smart Farming – A reviewSjaak Wolfert, Lan Ge, Cor Verdouw, Marc-Jeroen Bogaardt, Big Data in Smart Farming – A review, In Agricultural Systems, Volume 153, 2017, Pages 69-80, ISSN 0308-521X, https://doi.org/10.1016/j.agsy.2017.01.023.