World Journal of Engineering Research and Technology
WJERT
www.wjert.org
Impact Factor Value: 5.924

DATA ANALYTICS: RESEARCH ISSUES, CHALLENGES, TOOLS & ITS LIMITATIONS

[1]*Jogannagari Malla Reddy, [2]S.V.A.V. Prasad and [3]Kothuri Parashu Ramulu

[1]Prof & Head, Dept. of CSE Indur Institute of Engg. & Tech. Siddipet, Telangana State, India.

[2]Prof. & Dean Lingaya's Vidyapeeth Faridabad, Haryana, India.

[3]Assoc. Prof, Dept. of CSE Indur Institute of Engg. & Tech Siddipet, Telangana State, India.

Article Received on 23/01/2021 **Article Revised on 13/02/2021** **Article Accepted on 03/03/2021**

***Corresponding Author**
Jogannagari Malla Reddy
Prof & Head, Dept. of CSE Indur Institute of Engg. & Tech. Siddipet, Telangana State, India.

ABSTRACT

The modern digital technology generates the tremendous data with high velocity and stored in organization repositories. It is in large volume and more complex to process with traditional systems. Due to the rapid growth of such data, solutions are needed in order to handle and extract insights from these datasets. Big data analytics can solve such complexity. Big data and its analytics plays vital role in finding these insights for effective decision making. The analysis of the data can performs with robust tools such as Hadoop and framework like MapReduc etc. The basic purpose of this paper is comprehensive view to explore the potential impact of big data analytics challenges, research opportunities and associated tools of big data on the organizations.

KEYWORDS: Big Data, Analytics, Decision making, Hadoop, MapReduce, Competitive edge.

Data has become an ingredient of every industry, economy, organization and individual form long back. The rapid growth of digitization has led to several complimentary improvements in the industry that leads to large amount data collected from various sources which motivated the concept of big data. There is need to extract insights from the big data which full fill the business needs.

In recent years large volume of data has been generated from various domains like health care, marketing, finance, personal, transport information systems and other interdisciplinary areas. Apart from the information systems, the data accumulates from the digital technologies such as web documents, internet search indexing and social computing. Social computing refers social network analysis, recommender systems, reputation systems, online communities and predication markets.

The Big Data is collection of large and complex datasets that traditional data processing systems is inadequate to process it. Big data is structured, semi structured or unstructured large volume of data occupies to petabytes or exabytes.[1] The Big data analytics is the process of examining datasets and extracts into insights for decision making. The Big Analytics provides new opportunities for upcoming researchers in knowledge processing for decision making which is useful for organizations.

Today, Most of the organizations has motivated into a data-driven, they consists more data that is related to their customers, competitors, markets and business processes. This data stored, classified and analyzed to make sense as valuable nugget. Data Analytics is a set of quantitative and qualitative techniques for retrieve the valuable insights from the data. It encompasses many processes that include extracting data and classifying it in order to derive various patterns, relations and valuable insights.

Data Analytics have lot of demand in Enterprise Sector achieve the competitive advantage. Many tools have merged with various functionalities to fulfill the needs industry. These tools can convert data into meaningful insights would evidently be stand as leaders of the hyper-competitive world.

This research paper presents comprehensive overview of Data Analytics and its importance, applicability in the Business. The organization of the rest of the paper is as follows. The Section 2 describes the prior work related to the big data analytics. Section 3 states the overview spectrum & taxonomy of Big Data Analytics. The section 4 presents various tools involved in the Data Analytics. The research opportunities, Challenges and limitations highlighted in Section 5. Finally concluded with future scope in the Section 6.

2. Literature Review on Big Data Analytics

Recently, the eminent database theorists and researchers have worked on Big data analytics and its role in and decision making. The derived taxonomy can be useful in extend the knowledge discussion in the area of Big data analytics. The good number of researchers provided key research contributions in the area of data analytics is as follows.

- Roger Magoulas from 'O' Reilly media in 2005 was first introduced the term "Big Data" in order to define that the large amount data can not mange with traditional database system due its complexity and size.[5] Many research papers published in 2008 on Big Data Concept covering with implications in various fields.[5]
- According to Ed Dumbill Chair at the O' Reilly Strata Conference, described as "Volume of the data exceeds the processing capacity of Conventional database systems".[5]
- With reference to MIKE 2.0 is the open source Information Management, Big data is defined by its size, and complexity. It represents that it cannot be handled with traditional data base techniques due to inconsistency and unpredictability.
- Manish Kumar Kakhani and Seeti Kakhani,[8] highlighted the various research issues of Big Data analytics with its characteristics.
- Vivekananth.P et al,[12] compared and contrasted various data analytics methods used for social media analytics, text analytics, audio analytics, video analytics and predictive analysis.
- D.P Acharjya. et al,[4] conducted the survey on Big Data Analytics with its challenges, open Research Issues and Tools. The basic objective of the survey is to explore the potential impact of big data challenges, open research issues and associated tools, His survey open the new opportunities for the researchers.
- G. Sabarmathi et al,[6] proposed the research plan for health care system. She suggested the methodology to integrate the Data Analytics for effective health care modeling.
- K. Siddardha et al,[7] presented the paper on Big Data with importance in the industry and individual. The authors discussed various challenges & issues, limitations and tools involved in the big data analytics with more elaboration. It opens new horizons to innovators to develop the solutions, based on the challenges and issues.
- Aarushi Arya et al,[1] published the research paper on Big Data analytics and its applicability in the cyber security. The data analytics identify the patterns of activity that represent network threats. The Big data can improve the information security with best procedures.

- Nandhini. P,[10] described role of Big data analytics in the work information technology using with applications of Cloud Technology, Data Mining, Cloud Technology, Had0op and MapReduce. She analyzed the merits and demerits of various security methodologies proposed by different researchers.
- Mantrpatjit Kaur et.al,[9] discussed the role of Big Data Analytics in the area of Internet of Things(IOT). Large amount of data generated every day from modern digital technologies such as Internet of Things and Cloud computing which require lot of effort to extract knowledge for decision making. He identified the various challenges, open research issues, tools in perspective of Internet of Things.

3. Overview Taxonomy of The Big Data Analytics

Big data is structured, semi structured or unstructured huge quantity of data accumulated each day through modern information systems and web environment. The enormous amount of data is possible to extract insights with conventional systems. This massive dataset is available for decision makers in the form of Big data. The data can be mined to extract the information which is used in the machine learning applications, predictive modeling and other advanced analytics applications. Big Data Analytics is process of examining the data sets to extract the insights which is useful for the business systems.

The heterogeneous massive data is referred as "Big Data" which contains the following characteristics (5V's)

3.1 Characteristics

Volume*:* Volume refers the massive amount of data that are being generated from digital environment in every second. For example, Internet of Things with sensors all over the world generates the large volume of the data every second in the digital world. This amount of data in the world will be in the double in every two years. By 2025, that will be 75 times the amount of the data set what we had in 2011.

Variety: In the 1960, strongest data types are numbers and text only. Today, in addition to that other multimedia data is also used together. It needs different types of analysis and tools for processing. The Big Data may comprise the heterogeneous data, it provides summarization, lineage, privacy and auditability. But the Complexity arises when source of the data is changed without prior notice.[7]

Velocity: Velocity is the rate of growth and how fast the data gathered for analysis. The data velocity of big data is faster than conventional systems. The velocity of the data more such as phone conversation, data sent by sensors, Internet transmission data, stock exchange data in real time. High volume of data transmitted with rapid speed and needed to be analyzed.

Veracity: The huge amount of data generated from sensors is noisy and sometimes often incorrect. It refers confusion, disorderliness with lack of trustiness and accuracy are less controllable. Big data analytic technologies will help to work with this type of data.

Value: Unprocessed data is no value unless it is refined to obtain the structured information. Big data analytic technologies can derive the insights from massive data and predict the future trends to make suitable decisions.

3.2 Importance of the Big Data Analytics

Big Data analytics tackle the massive data and identify the new opportunities for the organizations. It leads to smarter business moves, more efficient operations, good profits, customer satisfaction. Data Analytics perform key role in improving the business.

Gathering hidden nuggets/ Insights: Hidden nuggets / insights retrieved from big data. The outcomes to be analyzed and inferences will be highlighted in the business system.

Report Generation: The reports generated in structured format and used by the respective teams and individuals for taking the effective actions and decisions.

Analytical Study of the Market Trends: With analytical Study, the pros and corns of the competitors can be find out. With the analytical study the business growth and profits to be improved.

Cost Reduction: Big Data Analytics technologies such as Hadoop and Cloud based application provides remarkable cost effective in storage for huge amount of data for business success.

Rapid & Effective Decision Skills: With robust tools of Data Analytics such as Hadoop analytics improve the decision making more faster and effective

Innovation of new products & services: The changing nature of technology, stakeholder behavior and market competition, the organization are moving towards the new definitions

and guidelines. With help of Data Analytics the business enterprises can able to analyze & predict the customer requirement & future needs. The outcomes of data analytics critical insights used for innovation of new products and services of the business. It leads to competitive advantage of Business organization.

4. Big Data Analytics Tools

In recent development, regardless with industry various robust analytical tools and technologies tools are available to process the big data. The tools used for parsing data or visualization which is used to making sense the data in effective manner. Digital enterprises such as Google, Amazon, Microsoft, Face book cannot survive without Data analytical tools. These organizations are hiring more number of Data Scientists.

In this section we concentrate on various Big data tools related to batch processing, stream processing and interactive analysis

Batch processing: Apache Hadoop and MapReduce Infrastructure such as Mahout and Dryad used for batch processing.

Apache Hadoop and MapReduce: The Apache Hadoop and MapReduce is robust software framework for the big data analysis. It consists various sub systems such as Hadoop kernel, MapReduce, Hadoop distributed File System (HDFS) and apache hive etc.[3]

The programming model of MapReduce is process the large datasets by using divide and conquer method. Method of divide and conquer implemented in two phases such as Map step and Reduce Step. Hadoop infrastructure works on two kinds of nodes such as master node and worker node. The master node decompose the input into smaller sub problems and distributes into worker nodes in map step. Later the master node combines the outputs for all the sub problems into reduce step. It provides fault-tolerant storage and high throughput data processing.

Apache Mahout: Apache Mahout is the commercial machine learning technique which is scalable for large scale and intelligent data analysis projects. Apache Mahout includes various algorithms such as classification, clustering, pattern mining, dimensionality reduction and regression, evolution algorithms, and batch based collaborative filtering executed on top of Hadoop platform with map reduce framework. The objective of mahout is to construct vibrant, responsive, diverse community to motivate discussion on big data projects and its use

cases. The reputed organizations such as Google, IBM, Face book, Amazon, and Twitter are using this scalable machine learning algorithms for their operations and overcome the challenges.[4]

Dryad: Dryad is popular programming model for executing the parallel and distributed programs for handling the large context base on dataflow graph. It provides the cluster of computing nodes and cluster of resources to the programs in a distributed pattern. Dryad provides thousands of machines and multiple processors or cores for user. The major advantage of the Dryad, the user need not know about the concurrent programming. This application executed a computational directed graph which composed with computational vertices & communicational channels. It provides various functionalities like job graph generation, schedule of machines for available processes, transition failure handling in the cluster, visualizing the job, collection of performance metrics and updating job graph dynamically in response to the decision policies without knowledge of semantics of vertices.[2]

Stream processing: Stream data processing tools used mostly on real time analytic application, large scale streaming platforms. Ex. Storm and Splunk.

Strom: Strom is fault tolerant distributed real time computation system for processing of huge stream data. It is unique designed real time processing in contrast with Hadoop of batch processing. It works similar to the hadoop cluster. The Strom cluster users run different topologies for different Strom tasks but Hadoop implements map reduce jobs for the applications. There are many variations between the both technologies. The difference is that map reduce job eventually finishes but topology process all kinds of messages of master node and worker node which pefforms Nimbus and Supervisor roles respectively. These two are similar roles of map reduce framework i.e, job tracker and task tracker. Nimbus distributes codes among the storm cluster and scheduling, assigning tasks to worker node. The Supervisor complies the tasks assigned by the Nimbus. The entire computations is decomposed and distributed to number of worker processes and each worker process implements as part of the topology.[4]

Splunk: Splunk is software tool used to retrieve and analyze the machine data which is the outcome of web applications, sensors, devices or any data created by user community. The tool analyze the log file which is in the form of structured and semi structured with proper

data modeling. Built-in features of the tools can recognize the data types, field separators, optimize the search processes and generate the visualization on search results. The product available three different categories such as Splunk Enterprise, Splunk Cloud and Splunk Light. The Enterprise Edition contains full fledge features such data integration, data indexing, Data searching, using alerts, dashboards and data model etc.

Interactive Analysis: The interactive analysis process permit the users to interact in real time environment with own analysis. The various big data platform such as Dremel, Apache Drill support the interactive analysis for developing the projects.

Dremel: Dremel is a scalable, interactive ad-hod query system for Big Data analysis which exceeds Hadoop capabilities combining with multilevel execution trees and columns data layout, it can aggregate trillion row tables in seconds.

Apache Drill: Apache drill is another distributed system bigdata tool for interactive analysis. It will support many types of data formats, query languages and data sources. It is specially developed for exploit nested data. Apache Drill scale up on 10,000 servers or more and reaches capability to process petabytes of data seconds.[2] Drill use HDFS for storage and map reduce to perform batch analysis.

5. Research Opportunities, Challenges and Its Limitations

Big Data Analytics is emerging research area in the data management. Many researchers published various research papers in different perspective on this domain to target the problems efficiently. It consists with various opportunities, challenges and limitations.

Research Opportunities

The extensive research is going on in the area of Big Data Analytics. But less no. of tools at existence. In research respective, any one move towards an innovative area of distributed Algorithm Development for better prospects. The other realms are Data visualization, machine learning, Statistical analysis are upcoming research areas in the field of Data Analytics.[8] Defends upon innovator, there is unlimited opportunities are there in this emerging field.

Big data analytics is universal application, which is applied in different industries and sectors to achieve the competitive edge. Some of the applications of Big Data in different sectors are transportation, health care, education, agriculture, social media and poverty.

Challenges

- Most of the organizations may not be data driven, They can't understand the advantages of the data analytics. Without knowledge they hesitate the features of data analytics.
- Big data is collections of heterogeneous dataset and affected by noise, sometimes inconsistent. So, efficient storing and processing are prerequisites for big data.
- Collection of data from different sources which leads to redundancy. Detecting and eliminating the redundancy may improve the storage capacity. The Big data storage effect not only cost and complexity in analytical processing. The robust tools with effective analytical techniques, data mining algorithms and machine learning algorithms may tackle this problems.[11]
- Big Data is now rapidly expanding in various domains such as physical science, biological, Biomedical and various engineering disciplines. There is lack of well trained professional as on exists which cannot full fill the industry needs.
- The management may not trust the outcomes/ insights of data analytics sometimes, it is difficult to understand how data can generate such insights.
- There should be proper coordination is needed in between the data analytics team members and user community at different phases of analytics such as scope definition, extraction and delivery of out comes.
- Big data heavily depend on the cloud environment, but it creates big data security risks. Particularly third-party applications of unknown, can easily impose the risks into the enterprise networks with less security standards.

Limitations: The big data Analytics had many benefits and also have certain limitations of the following.

- Accumulation of the Big Data is much faster than processing capability it reaches to burst in few years. It needs some new technology to work: otherwise we will get over run by data. Only the Cloud data centers can solve this problem.[7]
- Security is foremost for every technology. There is no exemption for Big Data from third party applications.
- The data collected from multiple sources and stored at single site. There may be a possibility of data inconsistency.

CONCLUSION

This paper focused on basic fundamentals, advantages of Big Data Analytics, the various tools involved in the Big Data Analytics. The various research issues, challenges highlighted to analyze this Big Data. Big data analytics is becoming crucial technology for automatic discovering of hidden patterns/ insights for organizations which is used for decision making, predicting and identifying new opportunities. It is understood that every Big Data platform has its individual focus. Different techniques will be used based on the data processing. It will provide guidance for extensive research in area of Big Data Analytics to solve the problems more effectively.

REFERENCES

1. Arushi Arya et al, "Big Data Analytics in Cyber Security", "International of Engineering Research & Technology, ICCCS, 2017.
2. Althaf Rahamana.Sk et al, "Challenging tools on Research Issues in Big Data Analytics", International Journal of Engineering Development and Research, 2018; 6(1).
3. Bindu. M.G.," A Survey on Big Data Analytics: Challenges and Opportunities", International Journal of Computer Science and Information Technologies", 2018; 9(4).
4. D.p. Acharjya et al, "A Survey on Big Analytics Challenges, Open Research Issues and Tools", International Journal of Advanced Computer Science and Application, 2016; 7(2).
5. Elen Geanina Alaru et al, "Perspectives on Big Data and Big Data Analytics", 2012; 3(4).
6. G. Sabarmathi et al," Big Data Analytics Research Opportunities and Challenges: A Review", International Journal of Advanced Research in Computer Science & Software Engineering, Oct, 2016; 6(10).
7. K. Siddardha et al, "Big Data Analytics: Challenges, Tools and Limitations", International journal of Engineering and Technical Research, Nov, 2016; 6(3).
8. Manish Kumar Kakhani et al, "Research Issues in Big Data Analytics", Internatinal Journal of Application or Innovation in Engineering & Management", Aug, 2013; 2(8).
9. Mantripatjit Kaur et al, "Big Data Analytics on IOT: Challenges, Open Research Issues and Tools", International Journal of Scientific Research in Computer Science & Engineering", June, 2018; 6(3).
10. Nandhini. P, "A Research on Big Data Analytics Security and Privacy in Cloud, Data Mining, Hadoop and Mapreduce", International Journal of Engineering Research and Applications", April, 2018; 8(4).

11. Priyanka Gautam, "Impact of Data Mining on Big Data Analytics: Challenges and Opportunities", Internatinal Journal of Computer Trends and Technology, Mar, 2018; 57(1).
12. Vivekananth. P. et al, "An Analutis of Big Data Analytics Techniques", International Journal of Engineering and Management Research", Oct, 2015; 5(5).

Author Profile

Jogannagari Malla Reddy obtained M.Tech(CSE) from JNTU, Hyderabad. and awarded with Ph.D(CSE) from Lingaya's University, Faridabad. At present working as Professor & Head, Dept. of Computer Science & Engineering, Indur Institute of Engg. & Technology, Siddipet which affiliated to JNTU Hyderabad, His area of specialization in Software Engineering, Object Oriented Analysis Design, Data Base Management Systems, Data Analytics and Management Information Systems. He published various research papers on information systems in reputed National and International Journals & Conferences.

Dr. S.V.A.V. Prasad awarded Ph.D from Andhra University, presently working as Professor & Dean(R&D) in Lingaya's University, Faridabad, has 30 years of experience in Teaching and R&D. He published various research papers in National and International reputed Journals & Conferences. He guided many students in research programmes in communication system and information system areas.

Kothuri Parashu Ramulu completed M.Tech(CSE) from JNTUH and awarded with Ph.D(Computer Science) from Rayalaseema University, Kurnool, Andhra Pradesh. At

present he is working as Assoc. Professor in Dept. of CSE, Indur Institute of Engineering & Technology, Siddipet which is affiliated to Jawaharlal Nehru Technological University, Hyderabad. His area of specialization in Database systems, Software Engineering and Programming Languages. He published various research papers in National and International journals in the area of computer science.