BIG DATA ANLYTICS: APPLICATIONS, CHALLENGES, TOOLS

[1]Kushwant Kaur,[2]Dr. Kanwalveer Singh Dhindsa,[3]Ishatpreet Kaur

[1]Assistant Professor
[1]Department of Computer Science & Engineering,
[1]Gulzar Group Of Institutes, Khanna, India

Abstract : This The term, Big Data, refers to a huge repository of terabytes of data that is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. It has been chosen as the most promising research trends of the decade, drawing attention from every segment of the market and society. Big data analytics process includes deployment and use of analytic tools, revealing hidden patterns, incomprehensible relationship and other important data to support enhanced decisions and gain competitive advantages over business rivals. The importance of big data is increasing day by day. The basic objective of this paper is to explore the application areas of Big data, potential challenges in BDA, open research issues, and various tools associated with it.

IndexTerms **- Big Data; Big Data Analytics; Big Data Applications;Tools; Challenges; Issues.**

I. INTRODUCTION

The term Big Data refers to large growing heterogeneous data sets thatis distinct in numerous ways such as volume (too big), veracity (much commotion), velocity (faster arrival), variability (quick changes), and variety (diversity).The BD is an assembly of information with unique excellence that for a problem realm at any given moment can't be expertly handled using current/accessible/apperceived/routine advancements and strategies that have a specific end goal to concentrate esteem.A new type of big data analytics, as well as different storage and analysis methods are required to handle such data sets that is large in the size, has variety, and changes rapidly. Such sheer amounts of big data need to be properly analyzed, and pertaining information should be extracted.

Big data sizes are constantly increasing. A single data set size ranges from a few dozen terabytes (TB) to many petabytes (PB). , The larger the set of data, the more difficult it becomes to manage. Consequently, some of the other difficulties related to big data include capturing, storing, searching, cleaning, integration complexities, sharing, analyzing, visualizing and lack of skilled personal. Nowadays, enterprises are exploring large volumes of highly detailed data to discover previously unknown facts. Big data analytics arean advancedtechniques that are applied on big data sets. Analytics based on large data samples reveals and leverages business change [10]. Analytics can be classified in to three types they are: Predictive Analytics, Descriptive Analytics and Prescriptive analytics. Advanced BDA requires extremely efficient, scalable and flexible technologies to efficiently manage substantial amounts of data – regardless of the type of data format (e.g. textual and multimedia content) [12].

As machines communicate with each other over data networks, the datafication concept and ever increasing technological advancements, advocates assert that in the future a majority of data will be generated and shared through machines[4]. All data available in the form of big data are not useful for analysis or decision making process. Industry and academia are interested in disseminating the findings of big data. This paper focuses on application areas, challenges in big data and its available tools and techniques.

II. BIG DATA ANALYTICS

Big data analytics is the process of examining large and varied data sets to uncover information including hidden patterns, unknown correlations, market trends and customer preferences that can help organizations make informed business decisions.Big data analytics is a form of advanced analytics, which involves complex applications with elements such as predictive models, statistical algorithms and what-if analysis powered by high-performance analytics systems.There are 4 types of analytics, which should be used for different types of data: [4]

i. **Predictive:**Predictive Analysis establish existing data patterns and gives list of solutions which may come for given situation predictive analysis study. This analytical method is one of the most commonly used methods used for sales lead scoring, social media and consumer relationship management data. Predictive Analysis can be applied for Weather forecast based on identifying the patterns from the history data[2].
ii. **Prescriptive:**Prescriptive analysis reveals actions and recommend of what step should be taken. It goes one step forward of predictive as it provides multiple actions with likely outcomes for each decision.
iii. **Diagnostic:**Diagnostic analysis is used to uncover any hidden patterns which help for complete root cause as well as identify any factors that are directly or indirectly causing effect. Diagnostic analysis is majorly used in social media for analyzing the number of posts, shares etc.
iv. **Descriptive:**Descriptive analysis also known as data mining operates what is happening in real-time. It is one of the simplest types of analytics as it converts big data into small bytes. The result is monitored through e-mails or dashboard. It is used by majority of organization.::

III. APPLICATIONS OF BIG DATA ANALYTICS

Big data analytics applications enable big data analysts, data scientists, predictive modelers, statisticians and other analytics professionals to analyze growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence (BI) and analytics programs. Some applications of Big data analytics are as follows:

i. **Smart Grid case:**In the real time, it is essential to manage the national electronic power consumption and to monitor Smart grids operations. Big Data analytics helps to identify transformers at risk and to detect abnormal behaviours of the connected devices[11]. Grid Utilities can thus choose the best treatment or action. The real-time analysis of the generated Big Data allow to model incident scenarios. This enables to establish strategic preventive plans in order to decrease the corrective costs. In addition, Energy-forecasting analytics help to better manage power demand load, to plan resources, and hence to maximize ports.

ii. **E-health:**The analysis of data sets, generated from different heterogeneous sources (e.g., laboratory and clinical data, patients symptoms uploaded from distant sensors, hospitals operations, pharmaceutical data) has many beneficial applications[13]. It enables to personalize health services to adapt public health plans according to population symptoms, disease evolution and other parameters.

iii. **Internet of Things (IoT):**The high variety of objects, increases the applications of IoT, which supports for logistic enterprises. It has become possible to track vehicles positions with sensors, wireless adapters, and GPS. Nowadays, companies not only supervise and manage employees but also optimize delivery routes[3]. This is by exploiting and combining various information including past driving experience.One of the best example of IOT application is Smart city.

iv. **Public utilities:**Public Utilities such as water supply organizations and Sewage Board are applyingBig data Analysis techniques. Bangalore Press has reported that these organizations are implementing a real-time monitoring system to detect leakages, illegal connections and remotely control valves to ensure equitable supply of water to different areas of the city. This application resultsin reducing the need for valve operators and timely identification and fixing water pipes that are leaking.

v. **Transportation and logistics:**RFID (Radiofrequency Identification) and GPS are used by many road transport companies to track busesand explore interesting data to improve there services. For example, data collected about the number of passengers using the buses in different routes are used to optimize bus routes and the frequency of trips. BDA also helps to improve travelling business by predicting demand about public or private networks. Prediction from such data is a complicated process because factors such as weekends, festivals, night train, starting or intermediate station has greater impact on it. Machine learning algorithms makes it possible to mine and apply advanced analytics on past and new big data collection.

vi. **Insurance companies:** Government for giving medical claim to patients do large amount of expenditure. By using BDA analysis, prediction and minimizing fraud medical claims is obtained[4].

vii. **Pharmaceuticals:** the techniques help R&D to produce drugs, instruments, tools etc in shorter period of time, which are effective in treating specific diseases[4].

IV. CHALLENGES IN BIG DATA ANALYTICS

Researchers and professionals are facing several challenges while exploring Big Data sets and when extracting value and knowledge. There are various levels of difficulties including: data capture, storage, searching, sharing, analysis, management and visualization. Also, there are security and privacy issues especially in distributed data driven applications. The size of Big Data keeps increasing exponentially but the current technological capacity to handle and explore Big Data sets, is only in the relatively lower levels of petabytes, exabytes and zettabytes of data. Some technological issues are as following:

i. **Big Data management:** Data scientists are facing the challenge of data collection, integration and storage, with less hardware and software requirements, tremendous data sets generated from distributed sources[3]. It is important to efficiently manage Big Data in order to facilitate the extraction of reliable insight and to optimize expenses. Big Data management that is the foundation of big data analytics,means to clean data for reliability, to aggregate data coming from different sources and to encode data for security and privacy. The goal of Big Data management is to ensure reliable data that is easily accessible, manageable, properly stored and secured.

ii. **Big Data cleaning:**To manage the complexity of Big Data nature is a big challenge(velocity, volume and variety) [7]. Also, its complex to process it in a distributed environment with a mix of applications. It is essential to verify the reliability of sources and data quality before engaging resources to ensure reliable analysis results. Its also challenging to clean data sets that may contain noises, errors or incomplete.

iii. **Big Data aggregation:**Another challenge is to synchronize outside data sources and distributed Big Data platforms (including applications, repositories, sensors, networks, etc.) with the internal infrastructures of an organization. Most of the time, it is not sufficient to analyze the data generated inside organizations. In order to extract valuable insight and knowledge, it is important to go a step further and to aggregate internal data with external data sources. External data could include third-party sources, information about market fluctuation, weather forecasting and traffic conditions, data from social networks, customers comments and citizen feedbacks. This can help, for instance, to maximize the strength of predictive models used for analytics.

iv. **Imbalanced systems capacities:**An important issue is related to the computer architecture and capacity. Although, the CPU performance is doubling using Moore's Law, and the performance of disk drives is also doubling at the same rate. However, the I/O operations do not follow the same performance pattern. (i.e., random I/O speeds have improved moderately while sequential I/O speeds increase with density slowly) [3]. Consequently, this imbalanced system capacities may slow accessing data and affects the performance and the scalability of Big Data applications.

V. BDA TOOLS

There arises a need for new tools and methods specialized for big data analytics, as well as the required architectures for storing and managing such data. Accordingly, the emergence of big data has an effect on everything from the data itself and its collection, to the processing, to the final extracted decisions. Consequently, [5] proposed the Big Data, Analytics, and Decisions (B-DAD) framework which incorporates the big data analytics tools and methods into the decision making process [5]. The framework maps the different big data storage, management, and processing tools, analytics tools and methods, and visualization and evaluation tools to the different phases of the decision making process. Hence, the changes associated with big data analytics are reflected in three main areas: big data storage and architecture, data and analytics processing, and, finally, the big data analyses which can be applied for knowledge discovery and informed decision making. Each area will be further discussed in this section. However, since big data is still evolving as an important field of research, and new findings and tools are constantly developing, this section is not exhaustive of all the possibilities, and focuses on providing a general idea, rather than a list of all potential opportunities and technologies.

i. **Apache Hadoop and MapReduce:**The most established software platform for big data analysis is Apache Hadoop and Mapreduce. It consists of hadoop kernel, mapreduce, hadoop distributed file system (HDFS) and apache hive etc. Map reduce is a programming model for processing large datasets is based on divide and conquer method. The divide and conquer method is implemented in two steps such as Map step and Reduce Step. Hadoop works on two kinds of nodes such as master node and worker node. The master node divides the input into smaller sub problems and then distributes them to worker nodes in map step. Thereafter the master node combines the outputs for all the subproblems in reduce step. Moreover, Hadoop and MapReduceworks as a powerful software framework for solving big data problems. It is also helpful in fault-tolerant storage and high throughput data processing.

ii. **Apache Mahout:**Apache mahout aims to provide scalable and commercial machine learning techniques for large scale and intelligent data analysis applications. Its algorithms includes clustering, classification, pattern mining, regression, dimensionalty reduction, evolutionary algorithms, and batch based collaborative filtering run on top of Hadoop platform through map reduce framework. The goal of mahout is to build a vibrant, responsive, diverse community to facilitate discussions on the project and potential use cases. The basic objective of Apache mahout is to provide a tool for elleviating big challenges. The different companies those who have implemented scalable machine learning algorithms are Google, IBM, Amazon, Yahoo, Twitter, and facebook[6].

iii. **Dryad:**It is another popular programming model for implementing parallel and distributed programs for handling large context bases on dataflow graph. It consists of a cluster of computing nodes, and user use the resources of a computer cluster to run their program in a distributed way. The major advantage is that users do not need to know anything about concurrent programming. Dryad provides a large number of functionality including generating of job graph, scheduling of the machines for the available processes, transition failure handling in the cluster, collection of performance metrics, visualizing the job, invokinguser defined policies and dynamically updating the job graph in response to these policy decisions without knowing the semantics of the vertices [9].

iv. **Storm:**Storm is a distributed and fault tolerant real time computation system for processing large streaming data. It is specially designed for real time processing in contrasts with hadoop which is for batch processing. It is also easy to set up and operate, scalable, fault-tolerant to provide competitive performances. The storm cluster is apparently similar to hadoop cluster.

v. **Apache Drill:**Apache drill is another distributed system for interactive analysis of big data. It supports many types of query languages, data formats, and data sources. It is designed to exploit nested data.

VI. CONCLUSION

There are many challenges in Big Data Analysis Process. Big Data platforms are supported by various processing, analytical tools as well as dynamic visualization. Such platforms enable to extract knowledge and value from complex dynamic environment. They also support decision making through recommendations and automatic detection of anomalies, abnormal behavior or new trends.

In this paper, the concept of Big Data Analysis has been discussed. In addition to that, the applications of Big Data in several domains has also been described. Besides, the paper focuses on the tools and technologies used in the process Big Data Analytics.

REFERENCES

[1] Acharjya D. P. 2016. A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools: International Journal of Advanced Computer Science and Application, 7(2): 511-518

[2] Bansal A. 2017,; Web Mining in E-commerce, International Journal of Advance Research and Innovation 2347 – 3258

[3] Chen, C.P., Zhang, C.-Y., 2014. Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. 275, 314–347

[4] Chunarkar-Patil, P.2018. Big data analytics: Open Access Journal of Science, 2(5): 326-335.

[5] Elgendy, N.2013: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, 164

[6] Ingersoll G. 2009, Introducing apache mahout: Scalable, commercial friendly machine learning for building intelligent applications, White Paper, IBM Developer Works, 1-18

[7] Khan, N., Yaqoob, I., Hashem, I.A.T., Inayat, Z., Mahmoud Ali, W.K., Alam, M., Shiraz, M., Gani, A., 2014. Big data: survey, technologies, opportunities, and challenges. Sci. World J.
[8]Li H., Fox G. and Qiu J. 2012, Performance model for parallel matrix multiplication with dryad: Dataflow graph runtime, Second International Conference on Cloud and Green Computing, 675-683.
[9] Nambiar, R., Bhardwaj, R., Sethi, A., Vargheese, R., 2013. A look at challenges and opportunities of Big Data analytics in healthcare. In: In: 2013 IEEE International Conference on Big Data. IEEE, 17–22.
[10] Russom, P.2011. Big Data Analytics. In: TDWI Best Practices Report,1–40
[11] Stimmel, C.L., 2014. Big Data Analytics Strategies for the Smart Grid. CRC Press.
[12] Sivarajah U. 2017. Critical analysis of Big Data challenges and analytical methods. Elsevier, journal of Business Research,70: 263-286
[13] Van Dijck J.2014, Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology Surveillance & Society, 12 (2):197-208