IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 6, Ver. IV (Nov.- Dec. 2017), PP 65-71
www.iosrjournals.org

An Avant-Garde Approach of Blockchain in Big Data Analytics

Archana Senapati [1], Pratyush Ranjan Mohapatra[2]
[1](*Department of Computer Science & Engineering, Gandhi Engineering College, India)*
[2](*Department of Computer Science & Engineering, Gandhi Institute For Technology, India)*

Abstract: *Administration of data and its analysis have been an ineludible part of the world for ages now. With the progression of big data, myriad number of companies that dealt with a large quantum of data gained momentum as major issues regarding management of data had found an optimistic solution. Even the field of analytics got a boost because of the numerous techniques introduced by big data for analysis of enormous quantity of data. Various methodologies introduced by big data for analytics that yield ground-breaking throughputs with high efficiency rates tend to develop complexities which could lead to damage of catastrophic scale. The major causes included security and storage which needed an immediate solution that is reliable and flexible. This gave rise to the concepts of decentralization and distributed system which when combined together engendered a new technology i.e. BlockChain. It is distributed ledger network that helps in making transactions without any centralized entity being needed. BlockChain is a full-proof solution to the problems of big data analytics as it not undergoes operations safely but also takes care of the storage issue. This newly developed technology has been in the light for a while now. BlockChain has found applications in various sectors which include industrial, medical, banking, and educational as well as defense. This paper discusses the concept of big data, its analytics and BlockChain. It elucidates the techniques and technologies involved in big data analytics and blockchain mechanism. It further discusses how Big data has impacted the canonical ways of handling data, the significance of Big data analytics and how the BlockChain Technology could be used similarly to tackle the issues in Big data analytics. The aim of this paper is to encourage further research in incorporating the BlockChain Technology into Big Data Analytics.*
Key Word*: Big data, big data analytics, Blockchain analgesia.*

I. Introduction

Data is nothing but a set of raw facts and figures. Data can be termed as a collection of values of a specific variable either qualitative or quantitative. It is a very important component in every research work. The data acquired is coupled with the pre-processing techniques which is further used massively in research [29]. Nowadays companies are flooded with data ranging from Petabyte to Exabyte to Zettabyte which gives birth to the most archaic problem of all eras i.e. storage and security. The main task of these ventures is to accumulate unstructured data from structured data, to dynamic from static paradigm shifts. It becomes serious to process such an amount of data and transform it into knowledge for decision-making activities. [25] This is when big data comes to light. Big data organizes and uproots the valued information from the fast growing, large volumes, variety forms, and frequently changing data sets collected from eclectic sources in a short span of time available using different statistical techniques [26]. Analysis of these data is majorly done in every field. Big Data analytics (BDA) is where advanced analytic techniques operate on big data. A paper in 2013 stated that BDA is a process of inspecting, cleaning, transforming, and modelling big data with the objective of discovering knowledge, generating solutions, and supporting decision-making [27]. But even big data had complexities which were clearly showcased in big data analytics mechanism. These problems to lead to hazardous situation wherein handling of dataset might seem quite similar to a nightmare. For instance, a social media platform like Facebook has to deal with tons of data, even a single error in analysis could result in huge amount of loss. The chief problem here is the dependence of every operation on a central entity. This was overcame by the arrival of blockchain in the market. A blockchain is a well-ordered collection of blocks, on which requires all the users to come to consensus. It arbitrates the history of asset control and provides a computationally unforgeable time ordering for transactions [28]. Blockchain not only provides storage and security but also gives a frugal plan of dealing with the data. It is widely inculcated in many streams.

II. Big Data

Big Data can be coined as a massive volume of both structured and unstructured data. It refers to huge datasets that are high in variety and velocity, which makes them difficult to handle using canonical tools and techniques [1]. Big Data begins with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data [2]. Four main features that characterize big data are: volume, variety, velocity, veracity.

- Volume: The volume of the data depicts the size of the data set. Normally it is measured in terabytes and gigabytes.
- Velocity: This defines the pace at which data is changing, or how often it is created.
- Variety: This refers to various the formats and types of data. Data obtained from numerous sources are of different types which are a part of the dataset.
- Veracity: It is related to the quality of data. Faster results could be obtained in analysis if data of good quality is provided.

The primary aim of Big Data applications is to help companies make considerable business decisions by incorporating advanced methodologies to analyze large volumes of data with less time consumption. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. Currently, Big Data is tremendously used in the field of analytics.

III. Big Data Analytics

Analytics is the process of examining data in to order to generate certain patterns from it which could be used to solve various aspects. Similarly, Big Data Analytics can be defined as the complex 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. BDA just cannot be called as a technology, it is an integral toolset of strategy, marketing, human resources and research. It analyses and mines big data to produce operational and business knowledge at an unprecedented scale and specificity [3].

Let us take an example to better understand this concept better. Here, we consider the product launch of Starbucks wherein it introduces new coffee product in the market.

- Starbucks was planning to introduce a new coffee product but was worried that customers would find its taste too mild. The day that coffee was rolled out, Starbucks monitored blogs, Twitter, and coffee forum discussion groups to assess customer's reviews. By noon, Starbucks discovered that although people liked the taste of the coffee, they thought the coffee was too expensive. Starbucks lowered the price, and by the end of the day all the customers were satisfied. Compare this fast response with a slow traditional way of waiting for the sales reports to come in and noticing negative impact on the sales are disappointing. The next step might be to discover the reason behind disappointing sales. And in several weeks Starbucks would have discovered the reason and responded by lowering the price.

Big data analytics offers various business benefits, including new revenue opportunities, more effective marketing, better customer service, improved operational efficiency and competitive advantages over rivals.

IV. Types of Big Data Analytics

It is beneficial to compare three kinds of analytics because the differences have implications for the technologies and architectures used for big data analytic. Performance of analytics varies depending on the platforms it is being executed.

- Descriptive Analytics: It is quite similar to a car's rear view mirror which is used to view vehicles at the backside or at a distance i.e. helps in analyzing things of past. It has applications in core area like reporting/OLAP, dashboards/scorecards, and data visualization [4].
- Predictive Analytics: This kind of analytics gives information about the future. Predictive Analytics make use of various methods and algorithms such as regression analysis, machine learning, and neural networks.
- Prescriptive Analytics: This technique gives suggestion while analyzing data. Prescriptive analytics provides optimal solutions for the allocation of scarce resources. It has been researched in academia for a long time but is now finding wider use in practice. To illustrate, the use of mathematical programming for revenue management is increasingly common for organizations that have "perishable" goods such as rental cars, hotel rooms, and airline seats [4].

As mentioned above big data is not only large, but also requires varied and fast-growing technologies and techniques to attempt extracting relevant information. Some of the most commonly used techniques and technologies are discussed below.

5.1. Techniques

- Association rule learning: It is a way of finding relationship among variables. Such method is often used in data mining and it also provides support to recommender systems like those employed by Netflix and Amazon [5].

- Data Mining: Data Mining is stated as "combining methods from statistics and machine learning with database management" by Manyika et al [6] in order to pinpoint patterns in large datasets. Picciano describes it as "searching or 'digging into' a data file for information to understand better a particular phenomenon" and files it among the most important terms related to data-driven decision making [7].
- Cluster Analysis: Manyika et al [6] describes cluster analysis as a type of data mining in which a large group is divided into smaller groups of similar objects whose characteristics of similarity are not known in advance and attempts are made to uncover the similarities among the smaller groups, and if they are new groups, what caused these qualities.
- Crowdsourcing: Basically, in Crowdsourcing data is being collected from a large group of people through an open call, usually via a Web2.0 tool. This tool is used more for collecting data than for analyzing it.
- Machine Learning: Machine Learning's main focus is to automatically learn to recognize complex patterns and make intelligent decisions based on data. An example given by Miller (2011/2012) of the U.S. Department of Homeland Security, which identified patterns in cell phone and email traffic, as well as credit card purchases and other sources surrounding security threats with the help of machine learning. These patterns were used to identify future threats so that they can be handled before they become large problems.
- Text Analytics: Text constitutes the major portion of generated data. Emails, internet searches, web pages, corporate documents, etc. are all based on text and can be useful as sources of information. Necessary information can be extracted by the use of text analytics from large amounts of textual data. This can be done to model topics, answer questions, and other goals.

Apart from this, other techniques used are classification, data fusion, network analysis, optimization, predictive modelling, regression, special analysis, time series analysis, and others.

5.2. Technology

Along with the analytical techniques, there are several software products and available technologies to facilitate big data analytics. Some of them will be discussed here:

- EDWs: Enterprise data warehouses are databases used in data analysis. For many businesses that are trying to start handling big data the big question is "Can the current or planned EDW handle big data and advanced analytics without degrading performance of other workloads for reporting and online analytical processing?" [8]. Most of the organizations manage their analytic data in the EDW itself while others use separate platform, which helps relieve some of the stress on the server resulting from managing your data on the EDW.
- Visualization Products: To find ways to visually represent results is one of the difficulties with big data analytics. Russom [8] lists this field as one of those having the most potential, saying it is "poised for aggressive adoption." Beyond simple representation, information search can be supported by visualization. An article from THE FOURTH PARADIGM (2009) [9] discussing in data-intensive science in which they explain that visualization products allow us to compare models and datasets and enables quantitative and qualitative decision-making. This article stresses scalability in visualization technologies and their ability to track provenance in real-time.
- Hadoop & MapReduce: MapReduce is a programming model generally used to support a lot of data simultaneously and Hadoop is one of the most popular open source applications of that model. Another article in THE FOURTH PARADIGM (2009) [9] written by Szalay and Blakeley in which they discuss and explain that the principles of MapReduce uses are similar to the "distributed grouping and aggregation capabilities that have existed in parallel relational database systems for some time" but they were able to scale very well to accommodate for exceptionally large data sets. Further they also explained that Hadoop implements a "data-crawling strategy over massively scaled-out, share-noting data partitions" where various nodes in the system are able to perform different parts of a query on different parts of the data simultaneously. The product works very well for big data, but for smaller projects it isn't as effective "when a good index might provide better performance by orders of magnitudes."
- NoSQL databases: These are the databases specially designed to deal with very large amounts of information that don't utilize a relational model. More Often, they are useful for tracking and analyzing real-time lists which grow quickly.

Big data analyst must evaluate their needs and choose the most appropriate tools for their company/organization. Changes are required, from business to business and also from sector to sector.

6. ISSUES OF BIG DATA ANALYTICS

Big data analytics is not a panacea for all problems in the field of analytics, and it raises some serious concerns that should be addressed and solved in a timely manner [11]. Major sectors contributing to challenges faced by BDA are:

- Privacy: It is the most cardinal issue with big data which includes conceptual, technical and legal significance [12]. Big data contains a lot of personal information about customers, clients, patients, and other types of users. [11] This information when combined with external large data sets leads to the inference of new facts. Information regarding the customer is collected and used to add value to the business without them being unaware of it [12]. The best paradigm to illustrate this issue would be social media apps which collect lots of personal information about an individual and after a point start suggesting unwanted stuff just to market an organization's products. This information given by an individual could be misused and could put the individual in jeopardy.
- Heterogeneity: Big Data consists of data sets that are collected from eclectic sources. This includes different formats, presentations and structures which tend to make the management of data a tedious process. Also, Variety, one of the key characteristics of big data leads to development of complexities in the analytics process. Data that cannot be managed is often termed as unwanted data [10]. Thus, proper measures should be taken to clean data. Various pre-processing algorithms are used to clean and scale data. Other Data Management tasks include transforming, clarification, dimension reduction and validation of data which are performed in the later stages [10].
- Timeliness: Nowadays applications like social network, sensor networks, semantic web and location based services are rapidly used by the customers on daily-basis. This requires the applications to process a variety of data constantly at faster rate [13]. But when the size of data set starts increasing it becomes difficult to process it. This leads to delay in the entire analysis process. Even a small-time gap could lead to heavy loss. Thus, time constraint should be kept in mind before undergoing any process.
- Capital: Computers have good storing capacities but when it comes to myriad numbers of data set even computers don't prove to be a proper solution. High amount money is required for storing and maintaining such data sets. Business ventures invest a lot of capital in getting proper resources for BDA. Apart from this, companies also invest in human capital to get expertise in the analysis of their data [11]

V. Block Chain

Blockchain is a chain of blocks carrying information that are connected to each other without having any central dominance. It can also be defined as a peer-to-peer digital ledger of transactions that may be either publicly or privately distributed to all users [14]. This technology thrives on the concept of decentralization and distributed system. Immutability is another cardinal characteristic of blockchain technology [16]. It means data could not be changed without approval from all the users. BlockChain is constantly growing as minors add new blocks to it to record the most recent transactions. These added blocks must contain information that are approved by peers in the chain [15]. This information stored in the blocks is then allowed to move freely. Any kind of update in data is done only after all the peers in the network give their consent.

Structure of Block

A block can be referred to a container which is used to store data. It is composed of header and body. The Block header consists of:

- Block version: determines the set of block validation rules to be followed.
- Parent block hash: It is a 256-bit hash value which points to the previous block.
- Merkle tree root hash: It is the hash value of all the transactions.
- Timestamp: indicates the current timestamp as seconds.
- N-Bits: gives current target of hashing in a compact format.
- Nonce: It is a 4-byte field that starts with 0 and increments for every hash calculation.

The block body further consists of a transaction counter and transactions. The maximum number of transactions that a block can hold depends on the size of the block and the size of each transaction. Blockchain makes use of an asymmetric cryptography to validate the authentication of transactions (NRI, 2015) [17].

MECHANISM

BlockChain mechanism is quite simple and secure compared to other technologies. The steps involved in this process are:

- Triggering a transaction: This is the initial stage of the mechanism. During this stage one entity begins to make a transaction by sending data i.e. a transaction gets triggered. This transaction is then broadcasted to all the peers in the network.

- Validation and verification of data: In this stage validation is done by the miners. The transaction broadcasted has a hash function attached to it which is used by the miners to get a proper output. Appropriate algorithms are chosen to generate accurate results. These results are then verified by every peer in the network. After approval from every node in the network it is passed to the next stage.
- Creation of new block: After successful validation and verification, formation of a new block takes places. This new block consists of Private key, hash function and the output generated in the previous step.
- Addition of block to the chain: The new block is then communicated to all the nodes in the network to be subsequently appended to the existing chain of blocks in the blockchain digital ledger [24].

Methodologies

Overall BlockChain depends on three methodologies namely Cryptographic key, Peer-to-peer Network and Protocol [18].

- Cryptographic key: It is an essential part of blockchain mechanism. It ensures that the source of transactions is legitimate and that hackers are unable to steal a user's funds [19]. Public key cryptography uses a pair of a public key and a private key to perform various tasks. Public keys are widely distributed among the peers, while private keys are kept hidden. A person willing to send data uses public key to encrypt it so that only the person with the private key can decrypt and read it. Private key is also used to generate a digital signature which helps other peers with the corresponding public key to verify that the message was created by the owner of the private key and was not altered since. This sort of authentication is chiefly useful when the size of data is large.
- Peer-to-peer network: Such a network aims to achieve the decentralization characteristics of blockchain. Updates or amendments could be made in data only after every peer in the networks agrees to proceed. This also helps in retaining authorization of the data which gets lost due to fraudulent methods.
- Protocol: It consists of set of operations that are to be performed in the mechanism. The Blockchain protocols keeps a record of whether the network works the way it was intended to by its creators, even though it's completely autonomous and isn't controlled by anyone. It also focuses on actions taking place while a transaction is being made i.e. sending/receiving of digital signature, nodes getting a copy of updated information etc.

Blockchain has plethora of benefits such as decentralization, persistency, anonymity and auditability. It is widely used in applications ranging from cryptocurrency, financial services, risk management, internet of things (IoT) to public and social services [17].

HOW BLOCKCHAIN HELPS IN BIG DATA ANALYTICS

Data and its analysis face a lot of challenges. And if the size of the data gets bigger the problem gets worse. One solution to this most common problem would be to inculcate BlockChain technology in data analytics. Blockchain incorporates a layer of security to the already secure Big data analytics technique making it more authenticated. This satisfies the two main Big Data Analysis demands: [20]

- BlockChain has network architecture that makes it nearly impossible to tamper the data by hackers and Trojans.
- It also allows operations to be performed on the data in a more reliable way.
- Blockchain ensures integrity, better performance, and more stability for managing data [21]. BlockChain technologies overcomes the challenges faced by big data for analysis. The key improvements observed using blockchain for big data analysis would be:
- Storage: Data is stored on individual nodes which are intelligently distributed with no central entity needing to control access to a user's files. This helps in improving security and decreasing costs via decentralized file storage [22]. For instance, the Blockstake naming storage system, has a four-tier architecture which fully utilizes the decentralization characteristics of the blockchain to ensure the high security of the data [21].
- Security: Data security is the primary benefit of using blockchain technology. Blockchain mechanism ensures that data is encrypted properly which makes modification of data a difficult task. Decentralization makes it easier to cross check file signatures across all the ledgers on all the nodes in the network and verify that they haven't been changed. It makes use of consensus protocols across a network of nodes, to validate transactions and record data in a manner that is incorruptible. Hacking of data seems to be impossible as all the data is not stored in a single repository. Exfiltration and corruption of data are unfeasible [23].
- Tools Accessibility: Lot of money is spent in getting tools for analytics. This problem could be solved by blockchain technology as it has managed to expand the availability of the tools with the help of

reorganization and democratizing the technology. Blockchain has allowed the big companies to try to make their analytics efforts more valuable and helpful for data scientists [20].

BLOCKCHAIN IN BANKING

BlockChain is widely used in banking and share market. There was an era when people had to stand in long queues to withdraw or deposit money in the bank. This traditional way of banking had a lot of drawbacks which include security issues and other things. But soon there was a solution to these problems which had inclusion of technology in it. Online transactions became a trend which every individual preferred. It reduced the time constraint which was one of the major concerns in the traditional banking ways, but this also lead to a new risk of money be stolen or account being hacked through online means. Apart from that, as everything turned online the complexities of managing these databases increased. All these problems led to incorporation of BlockChain technology in the banking industry. As BlockChain is a decentralized distributed ledger of transactions, it solved the chief problems of engendered by internet in a very easy and secure way. It helps to save the banking and transaction cost. Apart from providing faster and better services, it also helps to reduce the maintenance costs of the banks and share market. Fraud Detection is another advantage that is obtained by incorporating BlockChain. For instance, as a number of merchants go to multiple banks with the same invoice and get the bill discounted. It helps in tackling bad loans by getting through loan history [24].
Various banks like SouthIndian bank and ICICI have inculcated BlockChain in their system.

VI. Conclusion

This paper gives a brief about Big Data, its challenges in the field of Analytics and the role of BlockChain Technology to tackle those issues. The pertinence of these technologies and how they are connected is discussed ahead. The merits and demerits of Big Data Analytics are also emphasized along with the process of analysis. The aim of this paper is to encourage further research in assimilating BlockChain Technology for Big Data Analytics.

References

[1]. N. Elgendy, A. Elragal, Big Data Analytics: A Literature Review Paper, In the Proceedings of the 14th Industrial Conference, ICDM , 2014,Russia, pp.214-227
[2]. X. Wu, X. Zhu, G. Wu, W. Ding, Data mining with big data, IEEE Transactions on Knowledge and Data Engineering, 26(1), pp.97-107, 2014
[3]. Z.M. Bi, D.S. Cochran, Big data analytics with applications, Journal of Management analytics, 1(4), pp.249-265, 2014
[4]. H.J. Watson, Harnessing Customer Information for Strategic Advantage: Technical Challenges and Business Solutions, The Data Warehousing Institute, USA, 2014.
[5]. Chen, H., Chiang, R. H. L., & Storey, V. C., Business Intelligence andAnalytics: From Big Data to Big Impact, MIS Quarterly,36(4), pp.1165–1188, 2012
[6]. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H.,Big data: The next frontier for innovation, competition, and productivity,McKinsey Global Institute, USA, pp. 1–143, 2011
[7]. Picciano, A. G., The Evolution of Big Data and Learning Analytics in American Higher Education, Journal of Asynchronous Learning Networks, 16(3), pp. 9–20, 2012
[8]. P. Russom, TDWI Best Practices Report: Big Data Analytics, The Data Warehouse Institute (TDWI), pp. 16–35, USA, 2011
[9]. Hey, T., Tansley, S., & Tolle, K., The fourth paradigm data-intensive scientific discovery", Microsoft research, 2009.
[10]. M.A. Wani, S. Jabin, Big Data: Issues, Challenges and Techniques in Business Intelligence, In the proceedings of 50th Golden Jubilee Annual Convention, Springer, 2015
[11]. Dylan Maltby, Big Data Analytics, University of Austin, Texas, 2014
[12]. A. Katal, M. Wazid, R H Goudar, Big Data: Issues, Challenges, Tools and Good Practices, Sixth International Conference on Contemporary Computing, India, 2013, pp.404-409,
[13]. Shivaraj Koti, Shivananda V. Seeri, A Survey on Big Data Issues and Challenges, IOSR Journal of Computer Engineering ,19(2), pp. 75-78, 2017
[14]. JW MICHAEL, A COHN, JR BUTCHER, Blockchain Technology & Regulatory Investigations, Practical Law The Journal, pp.35-44, 2018
[15]. Melanie Swan, Blockchain: Blueprint for a New Economy, O'Reilly Media, pp. 40-75, 2015
[16]. F. Xavier, M. Zhegu, Research Handbook on Digital Transformation Edward Elgar, UK ,2016
[17]. Zibin Zheng, Shaoan Xie, Blockchain challenges and opportunities: a survey, International Journal Web and Grid Services, 14(4), 2018
[18]. Nolan Bauerle. How does blockchain technology work? Available at: [url+=https://www.coindesk.com/information/how-doesblockchain- technology-work/, 2018. Accessed Oct 2018].
[19]. Manisha Valera, Parth Patel and Shruti ChettiarT.S. Sharma, How does blockchain use public key cryptography? www.blockchaincouncil.org, 2018
[20]. How Blockchain Analytics find its way in Data Analysis May 23, 2018 /www.newgenapps.com/blog
[21]. Zheng BK, Zhu LH, Shen M et al., Scalable and privacy-preserving data sharing based on blockchain, JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 33(3), pp. 557–567, 2018
[22]. Blockchain Data Storage https://lisk.io/academy/blockchain-basics/use-cases/blockchain-data-storage
[23]. Here's How You Can Secure Your Data with Blockchain https://www.entrepreneur.com/article/318477
[24]. What is blockchain technology? Why it is believed to change the world? https://acodez.in/what-blockchain-technology/ Created: Aug 23, 2018

[25]. Lee, J., Lapira, E., Bagheri, B., & Kao, Hung-an, Recent advances and trends in predictive manufacturing systems in big data environment, Elsevier 2013,1(1), pp. 38–41.
[26]. R. Kune, P. K. Konugurthi, A. Agarwal, R.R. Chillarige and R. Buyya, The anatomy of big data computing, John Wiley & Sons, India, 2015
[27]. King, I, Lyu, M. R., & Yang, H., "Online learning for big data analytics", IEEE Big data, CA,2013
[28]. A. Back, M. Corallo, L. Dashjr, M. Friedenbach, G. Maxwell, A. Miller, A. Poelstra, J. Timón, and P. Wuille, Enabling Blockchain Innovations with Pegged Sidechains, 2014
[29]. Francisca Adoma Acheampong, Big Data, Machine Learning and the BlockChain Technology: An Overview, International Journal of Computer Applications (0975 - 8887) ,180 (20), 2018