REVIEW ARTICLE

Blockchain for Big Data: Approaches, Opportunities and Future Directions

Amrita Jyoti[1], Vikash Yadav[2,*], Ayushi Prakash[1], Sonu Kumar Jha[3] and Mayur Rahul[4]

[1]ABES Engineering College Ghaziabad, Uttar Pradesh, India; [2]Government Polytechnic Bighapur Unnao, Board of Technical Education, Uttar Pradesh, India; [3]Krishna Engineering College, Ghaziabad, Uttar Pradesh, India; [4]Department of Computer Applications, CSJM University, Kanpur, Uttar Pradesh, India

ARTICLE HISTORY

Received: November 13, 2022
Revised: March 05, 2023
Accepted: March 07, 2023

DOI:
10.2174/2352096516666230607104537

Abstract: The last several years have seen a significant increase in interest in big data across a range of scientific and engineering fields. Despite having several benefits and applications, big data still has some difficulties that must be overcome for a higher level of service, such as big data analytics, big data management, and big data privacy and security. Big data services and apps stand to greatly benefit from blockchain decentralisation and security features. In this article, we present an overview of blockchain for big data with an emphasis on current methods, possibilities, and upcoming trends. We begin by providing a succinct explanation of big data, blockchain, and the purpose of their integration. After that, we look at different types of blockchain assistance for big data, such as blockchain for security in big data collection, data privacy protection, storage, and collection. Next, we examine the latest work on the utilization of blockchain applications for big data across different industries, including smart grid apps and applications, smart city applications, and smart healthcare applications. A few illustrative blockchain-big data initiatives are given and discussed for a good understanding. Finally, difficulties and potential directions are examined to advance research in an exciting field.

Keywords: Big data, blockchain, blockchain services, distributed ledger, amazon, grid apps.

1. INTRODUCTION

Over the past ten years, data traffic has spread at an unmatched rate internationally, which is why "big data" has received so much attention. The big data business is expected to reach around 250 billion dollars before 2025, Big Data is a new generation technology being studied to evaluate vast amounts of data and identify its key features like analytics, knowledge discovery, and high velocity [1, 2]. Big data is regarded from a comparative perspective as datasets with very high dimensions and sizes that are unable to be managed, stored, analysed, or collected by present database methods. From an architectural point of view, big data is defined as a database with exceedingly high volumes, speeds, and representations that demand significant horizontal scaling techniques for effective processing [3].

Difficulties, privacy, and security are crucial concerns since big data usually incorporates many forms of sensitive information, like names, addresses, ages, and banking information. Many different techniques and solutions have been researched for protecting data privacy and confidentiality, and reinforcement learning's potential applicability was examined in [4]. To improve the data standards and handle the computationally in-depth activities required by IoT tools while providing security and privacy guarantees [5]. Blockchain can drastically change the way that present big data systems are run. In this survey, we look at blockchain for big data from all angles, including how it works, what it can do, and what the future holds for it.

1.1. The State of the Arts Today and our Contributions

There have been several reviews released in related fields over the past few decades due to the significance of big data and blockchain. One of the very first studies on blockchain was conducted to examine the security and privacy concerns with blockchain applications [6-9]. The survey presented game theory applications for blockchain systems, such as games for managing mining operations, games for addressing security and privacy concerns, and games for blockchain applications [10]. The use of blockchain in various technologies has been studied through several surveys. For instance, the potential use of blockchain for IoT systems in addition to related issues (including security, scalability, and data management) discussed [11-14]. The incorporation of blockchain with 5G and edge computing platforms was investigated ~~[15, 20]~~. Reviews of blockchain's uses and potential for smart

*Address correspondence to this author at the Government Polytechnic Bighapur Unnao, Board of Technical Education, Uttar Pradesh, India;
E-mail: vikas.yadav.cs@gmail.com

2352-0965/xx $65.00+.00

grid networks were conducted for the polls [17, 18]. Recently, surveyed the uses of blockchain in the present healthcare system [19]. A thorough analysis of the interoperability of blockchain technology is provided [20]. A thorough analysis of several experimental techniques, analytical models, and theoretical blockchain modelling was published in other intriguing studies [21, 22].

The concepts and applications of big data analytics have also been the subject of numerous surveys and given a survey of methods and tools for large data management [23]. Representative surveys may be found and big data analytics have also been used in intelligent transportation and smart grid systems [24, 25]. In conclusion, blockchain may enhance big data by improving data integrity, security, and privacy, enabling real-time data analytics, enhancing data sharing, and enhancing big data quality.

The blockchain application in developing private and secure drone big data solutions was highlighted in a recent article [26]. This review also put forth a secure plan for big data from drones that were built on a four-layer architecture comprising layers for users, data, clouds, and blockchain. However, this study only discusses blockchain drones for big data and is unable to go into great detail about the function of blockchain for big data. Similar to our paper, other surveys address the interaction between big data and blockchain but they only offer cursory overviews rather than an in-depth survey [12, 16, 27].

We specifically aim to present a thorough assessment of blockchain for big data, covering foundational information, contemporary methodologies, prospects, research problems, issues, and future perspectives. This review’s main objective is to examine the most recent research and conduct an assessment of how well blockchain fits into applications of big data.

We demonstrate how blockchain holds enormous promise for advancing big data analytics, including the administration of data sharing, improved security and privacy, control over unclean data, and improved data quality.

2. AN OVERVIEW OF BLOCKCHAIN AND BIG DATA

This section provides background information on blockchain, big data, and the drivers behind their integration.

2.1. Blockchain

To put it simply, the blockchain is a "distributed ledger" of comparable data items, or "blocks," that are linked together. Each block of this ledger is connected through cryptography, and it is continuously growing. A blockchain's stored data is a shared database that is regularly updated. The fact that this database is not stored or gathered in one place is one of the powerful advantages of the blockchain that contributes to its high level of security. Because there are a few duplicate copies of the ledger and a lot of personal computers (PCs) on the network, it would require a lot of processing power to get into the network and corrupt the records. Although it is theoretically conceivable to quantify the amount of computing power required to carry out a hack, this is practically impossible [28].

The blockchain's "blocks" are made up of PC code that contains information and can be altered to represent anything from money to a birth certificate. Through secure encryption, every "block" is linked to other blocks, creating the "chain." This "chain" can be compared to any representation of a conventional database because it combines data. When viewed as a whole, the blockchain can be compared to an accounting ledger, which contains a record of transactions. The blockchain's programming relies on a "distributed ledger" rather than keeping track of transactions on a local ledger. In practical terms, the distributed ledger is also known as the blockchain. A synchronised database that is stored simultaneously on thousands of PCs all around the world is the distributed ledger. The ledger is widely distributed and has had numerous undetectable copies at various times [29]. To grasp the impact of this invention, one must understand distributed ledger rules.

Furthermore, as stated, "Blockchains are distributed digital ledgers that record, authenticate, and prevent duplication of transactions using algorithms or an exact set of instructions without the need for a central authority," a more thorough definition of the blockchain is provided. This definition provides a more in-depth look at the blockchain by highlighting its key characteristics and reiterating that it is both a distributed technology and a decentralised system.

2.1.1. Constructional Elements of Block Chain Frameworks

Frameworks can be built with the help of blockchain technology [30]. The primary building blocks of blockchain frameworks are depicted in Fig. (**1**).

2.1.2. Data source

A database is a type of information structure used to store data. By fusing information from several databases, it uses a social model to provide increasingly composite techniques for querying and information collection. Using a database management system (DBMS), the stored data can be sorted. The database is one of the core elements of the blockchain. Considering that this is unquestionably not your ordinary database with lines and parts, Instead, it is a record of all previous transactions for each participant client in the blockchain network. High-throughput, decentralised control, changeless information storage, and implicit security are characteristics of this type of database.

2.1.3. Miner

A miner is a central processing unit (CPU) that tries to solve a computationally challenging numerical problem in search of a new block [31]. To try to find the answer to the numerical puzzle, the miners can either work alone or in groups called pools. By informing all users of the blockchain network of new connections, the process of discovering another block is started. As the minor who discovers the block purchases the expenditures or charges of all the transactions in that block, in a few instances, the transactions with the highest expenses are first selected from minors.

2.1.4. Mechanism for Consensus

Trust frameworks involve using the strength of the scheme to monitor transactions. These are the foundations of

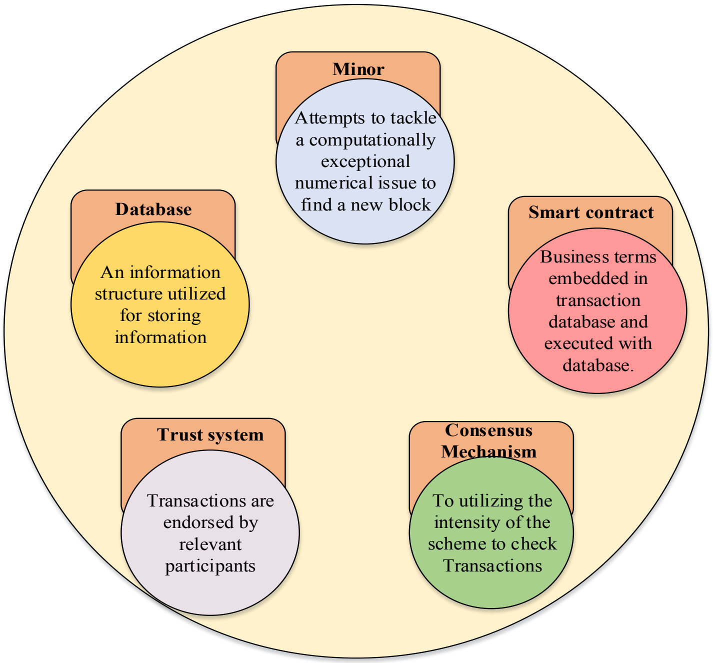
Minor
Attempts to tackle a computationally exceptional numerical issue to find a new block
Database
An information structure utilized for storing information
Smart contract
Business terms embedded in transaction database and executed with database.
Trust system
Transactions are endorsed by relevant participants
Consensus Mechanism
To utilizing the intensity of the scheme to check Transactions

Fig. (1). Blockchain building blocks.

the blockchain network scheme to monitor transactions. These are the foundations of blockchain networks. They provide the basis of blockchain applications, and since not all approval is obtained by consensus, we accept that the term "trust framework" is preferred over that of "consensus mechanism." This fundamental tenet of trust guides the shared goals and enthusiasm in blockchain architecture. The trust architecture changes with each new player in the blockchain market, creating variants that are specific to the stated blockchain use cases [32]. The three pillars of blockchain innovation are trust, exchange, and possession. The trust framework facilitates transactions for trade between cooperating organisations and exchanges between organisations. The ideal trust framework for explicit use cases, such as P2P and sharing economy models with several models, still need to be characterised.

2.1.5. *The Smart Contract*

Smart contracts are PC protocols that support, validate, or carry out the arrangement or execution of an agreement or that obviate the need for a formal declaration. Most smart contracts also have a user interface and frequently mimic the logic of authoritative provisions. Many different types of legally enforceable clauses can be made partially or entirely self-executing, self-authorising, or both, according to smart contract proponents. Smart contracts aim to provide security that is superior to that conventional smart contracts and to reduce other contracting-related exchange expenses.

2.2. BIG DATA

The background detail on big data's primary properties, uncertainty, and the analytics techniques that address big data's intrinsic uncertainty is covered in this part.

Big data was mentioned to be the upcoming frontier for innovation, competition and productivity in ~~May 2011 [36].~~ Over 3.7 billion individuals used the Internet in 2018-a 7.5% increase from 2016 [33-37]. The quantity of data formed globally increased from 1 zettabyte (ZB) in 2010 to 7 ZB in 2014 [38]. Variety, Velocity, and Volume (the three Vs) were found to recognise the developing features of big data in 2001 [39]. The four Vs-Value, Velocity, Variety, and Volume-were also used by IDC to define big data in 2011. Veracity was added as the fifth attribute of big data in 2012 [40-42]. Although there are numerous other V's, we focus on the five traits of large data, which are shown in Fig. (**2**).

Volume characterises the scope and extent of a database and alludes to the huge quantity of information generated per second. Because the amount of data and its type might affect how it is defined, it is difficult to establish a common criterion for large data volumes [43]. Currently, exabyte (EB) or ZB-sized datasets are typically regarded as big data, although problems remain for small-sized datasets [35, 44]. For example, every hour, Walmart collects 2.5 PB from around a million shoppers [45]. When attempting to study and comprehend the data and information at scale, many of the currently used data analysis tools may fail since they aren't intended for massive datasets [35, 45].

Variety mention to the various data types that are likely to be found in a database, such as multimedia and text information, which is random and challenging to analyse, but structured data, such as that kept in a relational database, is typically ordered and can be quickly sorted out. The semi-structured data contains tags to divide data items (such as in NoSQL databases), but it is up to the database user to enforce this structure [43, 46]. The presentation of mixed types of data, switching between different data types (like from unordered to ordered data), and conversion to the database's key structure during run time may cause uncertainty. Traditional big data analytics algorithms confront difficulties when managing multi-modal, imperfect, and noisy data from

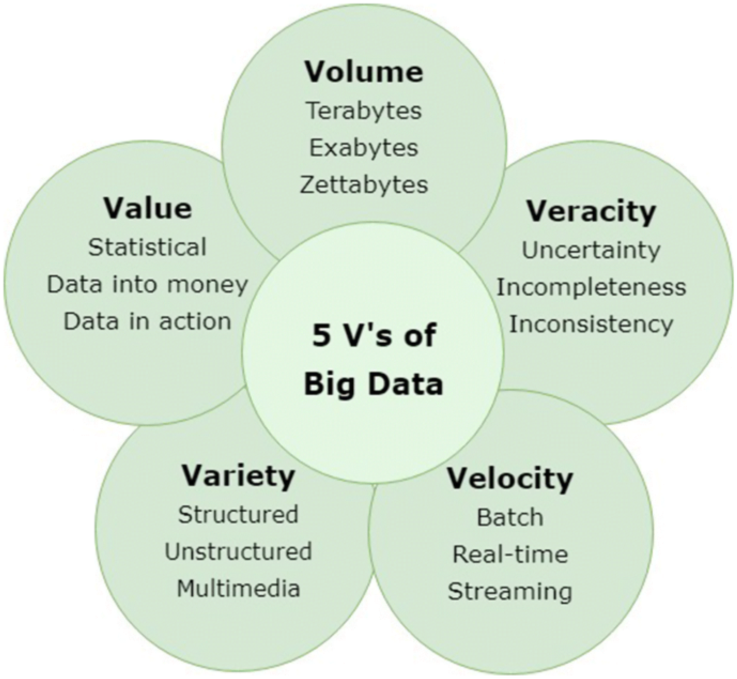

Fig. (2). Common big data characteristics.

the perspective of variety. Such techniques (such as data mining methods) can't take well-formatted input data [34]. Although the dataset itself can also be affected by uncertainty, this paper targets how uncertainty affects big data analysis.

It might be hard to analyse semi-structured and unstructured data effectively since they originate from sources with a diversity of data kinds and representations. Real-world datasets, for instance, are adversely affected by noisy, conflicting, and insufficient data. To eliminate noise from data, different pre-processing methods, such as data manipulation, data integration, and data cleaning, are utilised. Techniques for cleaning data address issues with data uncertainty and quality created by variance in large data sets (*e.g.*, inconsistent and noisy data). These methods for minimising distracting elements from the analysis method can considerably improve the effectiveness of the analysis of the data.

Velocity is defined as the rate of data processing (expressed in respect of batch, streaming, real-time, or near-real time), show up the demand that the rate of processing of the data match with the production rate of the data [8]. For instance, IoT gadgets constantly produce a lot of sensor information. A pacemaker which reports emergencies to a facility or doctor is an example of a gadget that monitors medical information and may cause patient damage or death if data processing and delivering the results to physicians is delayed [40]. Similar to how real-time OS enforces rigorous execution timing rules on devices in the cyber-physical world, big data applications may run into issues when their data is not supplied on time.

The level of data quality is represented by veracity (*e.g.*, imprecise or uncertain data). For instance, according to IBM's estimate, impoverished data standards cost the US financial system $3.1 trillion each year [41]. Veracity is categorized into three types: undefinable, terrible, and good, terrible, since data may be noisy, inconsistent, unclear, or incomplete. Trust and accuracy are difficult to start in big data analysis because of the amount and variety of data sources. For example, a worker may occasionally use an identical account to post personal ideas on Twitter while still utilizing it to publish official company information, making it hard for any techniques built to leverage the Twitter database to function properly.

Far from the preceding V's, which focused on indicating the problems in big data, the value represents the utility and context of data for a conclusion. For example, Amazon, Google, and Facebook have applied analysis to enhance the quality of big data in their products. To give product suggestions and boost customer satisfaction and sales, Amazon inspects a huge database of customers and their purchases. To improve its position in Google Maps, Google gathers location data from Android customers. Facebook keeps an eye on user interest to produce friend-tailored ads and recommendations. These organisations have grown remarkably as an outcome of data analysis, analysing huge amounts of information and gaining and reclaiming insightful knowledge that helps them make wiser business decisions [47].

2.3. Motivations for Integrating Big Data with Blockchain

The following is a discussion of the reasons for merging blockchain with big data.

Enhancing Big Data Privacy and Security: As the quantity of devices attached to the Internet rises daily, so does the volume of data being kept on external sites like the cloud. This produces new problems, like data attacks or breaches from interested outsiders [48]. Because of the reality that big data is not stored within a company's web area, existing privacy solutions such as firewalls aren't able to tackle this is-

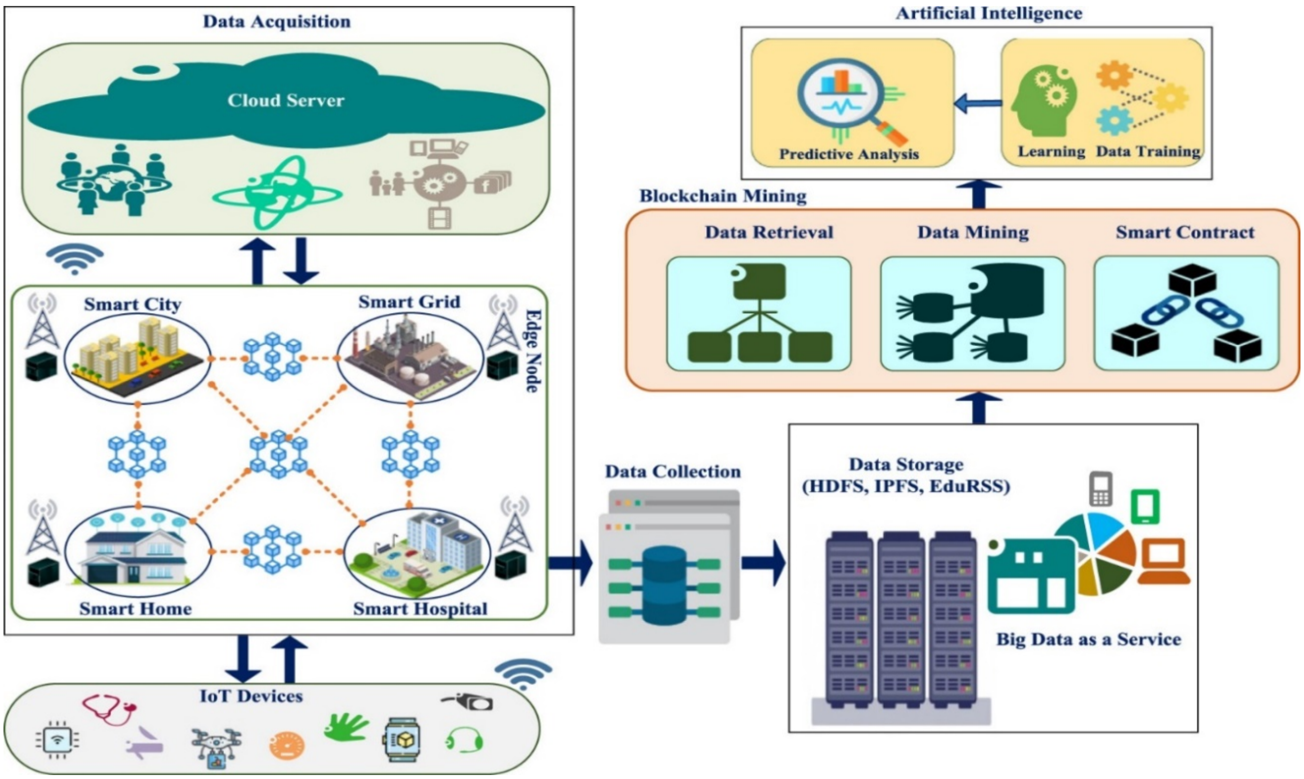

Fig. (3). Big data environment: blockchain services. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

sue. This is by enterprises that don't have authority over the information. Blockchain, a warehouse of huge amounts of information, can resolve this issue. It is exceedingly challenging for any unwanted access to the data due to the data's encrypted and decentralised storage in the blockchain network.

Enhancing Data Integrity: There is a possibility that somebody will interfere with big data information to affect the analysis forecast. It is virtually not possible to interfere with the information acquired through the blockchain connection thanks to the immutability feature of the blockchain. It is nearly possible to render the information in blockchain since doing so will require replacing it in not less than half of the network's nodes. Moreover, the blockchain's stability characteristic guarantees the correctness of the information stored there.

Real-time data analytics is now possible thanks to the blockchain's ability to store each transaction. Financial and banking organisations may resolve cross-border business, plus sizable sums, by enclosing real-time credit to a blockchain combination with big data analysis. Financial organisations may also look at data interchange in actual time, which gives them the ability to respond immediately to block transactions, for example.

Improvement in Data Sharing: By combining blockchain technology with big data, business providers can share information with stakeholders while minimising the risk of information leakage. Moreover, since each observation is recorded on the given blockchain, the inspection of the massive amount of data stored from different origins does not require to be redone.

Enhancement of Big Data Quality: Data technologists consume a lot of time integrating data because different origins use various formats when collecting data. Since the data is structured and comprehensive when stored on a blockchain, its quality can be improved. As a result, information scientists can use high-quality information to produce more precise forecasts in real-time.

3. BLOCKCHAIN APPLICATIONS FOR BIG DATA

Blockchain applications for big data are covered in this area, as shown in Fig. (**3**).

3.1. Using Blockchain to Acquire Huge data Securely

Big data applications are becoming more popular nowadays, but they still confront significant security challenges. In the process of processing data, gathering data is a crucial step. Untrustworthy data sources and communication channels make data collection vulnerable to different dangers and hostile attacks. As a result, safe data collection techniques are essential for many data applications. To date, some research projects have been conducted to offer secure data collection. For mobile crowd sensing (MCS), for instance, a safe massive data collection method based on blockchain is presented [49]. As a result of a grouping of cloud servers and MTs, an MCS system is formed. The MCS server records various sensing-based tasks and picks MTs in the region to execute them. Finite energy resources in MTs, the variety of sensing equipment, and securing the sharing of data across MTs are the biggest problems with data acquisition.