Cluster Computing (2024) 27:5547–5570
https://doi.org/10.1007/s10586-023-04257-7

Scalability of blockchain: a comprehensive review and future research direction

Iqra Sadia Rao[1] · M. L. Mat Kiah[1] · M. Muzaffar Hameed[2] · Zain Anwer Memon[3]

Received: 3 September 2023 / Revised: 20 November 2023 / Accepted: 21 December 2023 / Published online: 16 February 2024

Abstract

This comprehensive review paper examines the challenges faced by blockchain technology in terms of scalability and proposes potential solutions and future research directions. Scalability poses a significant hurdle for Bitcoin and Ethereum, manifesting as low throughput, extended transaction delays, and excessive energy consumption, thereby compromising efficiency. The current state of blockchain scalability is analyzed, encompassing the limitations of existing solutions such as Sharding and off-chain scaling. Various proposed remedies, including layer 2 scaling solutions, consensus mechanisms, and alternative approaches, are investigated. The paper also explores the impact of scalability on diverse blockchain applications and identifies potential future research directions by integrating data science techniques with blockchain technology. Notably, nearly 110 primary research papers from reputable scientific databases like Scopus, IEEE Explore, Science Direct, and Web of Science were reviewed, demonstrating scalability in blockchain comprising several elements. Transaction throughput and network latency emerge as the most prominent concerns. Consequently, this review offers future research avenues to address scalability challenges by leveraging data science techniques like distributed computing and parallel processing to divide and process vast datasets across multiple machines. The synergy between data science and blockchain holds promise as an optimal solution. Overall, this up-to-date understanding of blockchain scalability is invaluable to researchers, practitioners, and policy makers engaged in this domain.

Keywords Blockchain · Data science · Apache Kafka · Scalability · Machine learning

1 Introduction

Blockchain, initially surfaced in 2008, is a distributed ledger technology that provides a secure and transparent way to record and verify transactions [1] . It allows users to store and exchange data without the need for central authority [2], making it a decentralized system. In blockchain, each transaction is verified by a network of participants, who use cryptographic algorithms to validate the transaction [2, 3] and add it to the existing chain of blocks.The first blockchain, Bitcoin, was introduced in 2009 [2, 4]. It was designed as a peer-to-peer electronic cash system [2] that enables secure and anonymous transactions without the need for intermediaries. Ethereum, a blockchain platform introduced in 2015, is a well known implementation of Blockchain technology [1, 2]. It is designed to be more than just a digital currency e.g., Bitcoin. Ethereum is a decentralized platform that enables developers to build decentralized applications (dApps)

✉ Iqra Sadia Rao
iqra@um.edu.my

✉ M. L. Mat Kiah
misslaiha@um.edu.my

M. Muzaffar Hameed
muzaffar@bzu.edu.pk

Zain Anwer Memon
zain.memon@usindh.edu.pk

[1] Department of Computer System & Technology, Universiti Malaya, Malaysia, Kaula Lampur 50603, Malaysia

[2] Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan

[3] Department of Electronics Engineering, University of Sindh, Jamshoro, Pakistan

using smart contracts. Smart contracts are self executing programs that run on the Ethereum network and allow developers to create complex applications that can automate tasks, verify the authenticity of data, and enforce the rules of the system. While Blockchain technology in its current form has become particularly active study subject in information and communication technology (ICT). The goal of Blockchain technology is to establish transactional order in a distributed ledger without depending on trusted third parties [2]. However, to achieve this goal involves dealing with various issues, such as transaction rejection by miners and the high transaction fees dilemma [2, 5]. The former issue is self explanatory and may be solved by involving the use of digital signatures, the latter is caused by utilizing a single digital token to pay different entities, which changes the transaction history following a payment. Although, the emergence of decentralized solutions provides personal proof of their viability [6–8], experts are skeptical about it scalability. Scalability refers to a system's ability to accommodate exponential consumption growth while maintaining its functionality. The scalability issue with blockchain emerges when the number of nodes and transactions increases. This problem exists in major public blockchain systems, specifically Ethereum [2, 9–11], since each node must store and perform a computational function and each transaction must be verified.

Blockchain thus requires a lot of computing power, fast internet connection, and a lot of storage space all the time. The two most contentious blockchain performance metrics, transaction throughput and transaction latency, have yet to reach a suitable Quality of Service (QoS) level in many current and widely used public blockchain systems [3, 4, 3–4]. For instance, Ethereum can manage *12* transactions per second (TPS) [3]. Several researchers discussed the topic of the trilemma, involving decentralization, scalability, and security, was first described by Vitalik Buterin, the cofounder of Ethereum [5] and claims that these are the key parameters that must be traded off (see Fig. 1). Where, Decentralization is central to the concept of blockchain [2, 6, 8–13, 8–13], Security is a must have feature, while scalability remains the key complexity that must be addressed. In other terms, the blockchain trilemma asserts that trade-offs between these blockchain properties are almost unavoidable [71, 72].

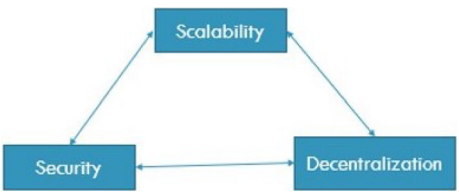


Fig. 1 Blockchain trilemma

In this paper, we applied Systematic Literature Review (SLR) to the most recent studies on the scalability frameworks on the Blockchain using Machine Learning and Data Science Techniques [15, 16, 58, 73]. Machine learning and data science techniques are introduced into the blockchain field, especially for scalability, because they offer various solutions to the scalability challenges faced by blockchain technology. Data science techniques such as distributed computing and parallel processing [58, 73] can be used to divide and process large datasets across multiple machines, reducing the computational requirements [12, 14], and increased processing speed of blockchain transactions. By combining blockchain and data science techniques, it is possible to improve the security and privacy of blockchain transactions. Data science techniques [16] can be used to detect and prevent fraud, as well as to identify potential vulnerabilities in blockchain systems. Machine learning algorithms can be used to optimize the consensus mechanisms of blockchain, improving the speed and efficiency of transaction [54]. Furthermore, introducing machine learning and data science techniques into the blockchain field can help to overcome scalability challenges, optimize consensus mechanisms, improve transaction processing speed, and enhance the security [25] and privacy of blockchain transactions.

The blockchain consensus mechanism was carefully analyzed and connected to all scaling-related components or factors [34]. Many academics have sought to address this issue in either an on-chain or off-chain fashion. On-chain options include, block size increasing, sharding, consensus procedures and sagwit [2, 4, 35]. A lighting network, on the other hand, is off-chain. This SLR focuses on the Ethereum Blockchain solutions addressed with Machine Learning and data science techniques [59, 74–76] and there is still a lot of work that needs to be done in this area since very less relevant content was found.

This SLR was initiated by first conducting a thorough research of the scalability issue in major public and private blockchain technology applications to identify the impacts of blockchain technology implementation in various sectors. After that, potential factors linked with transaction throughput, network latency, block size, and other issues were explored and tracked [13–29, 50, 13–29, 60, 65, 13–29].

The unique contribution of this paper is mentioned below:

- Presents a comprehensive review of the scalability challenges facing blockchain technology and identifies potential solutions and future research directions.

- The paper examines the current state of blockchain scalability, including the limitations of existing solutions, and investigates various proposed solutions, and other proposed methods.
- It explores the impact of scalability on various blockchain applications and identifies potential future research directions in this field combining Data Science techniques and blockchain technology since it has not been combined into a SLR before.
- Focused on combining blockchain and data science techniques, it is possible to improve the scalability, security and privacy of blockchain technology.
- Additionally, the paper reviews 110 primary research papers from renowned scientific databases to provide an up-to-date understanding.

This systematic review article is divided into several sections. Section 2 presents motivation in performing this study, Sect. 3 elaborates on blockchain, its components, characteristics and challanges. Section 4 emphasizes data science and its integration in the blockchain technology, while Sect. 5 discusses relevant work published by the research community. SLR methodology is presented in Sect. 6 and research questions and existing solutions with ML and Data Science solutions are presented in Sect. 7. The outcomes of this study are given in Sect. 8 with Open issues and future research direction are illustrated in Sect. 9, while the study is concluded in Sect. 10.

2 Motivation

The primary concern in performing this study arose when authors started working on limitations of blockchain technology, specifically scalability. This interest resulted in brainstorming about various solutions to improve scalability by combining data science and machine learning approaching with blockchain. The authors found very less literature regarding blockchain scalability with data science techniques, like distributed computing and parallel processing. Current research focuses, in general, the improvement in scalability parameters with traditional approaches which had trade-off between privacy and security. Security is crucial in blockchain, as it involves the financial transactions. Therefore, researchers face limitations dealing with scalability as well as security. Moreover, most solutions are complex and energy-hungry, and are challenging for researchers to analyze and develop blockchain solutions. This lack of research makes it difficult to identify relevant papers and develop effective solutions. The combination of Data Science and Blockchain is identified as promising direction in dealing with the blockchain scalability. This SLR, thus, focuses on multiple studies related to scalability with data science approaches, to highlight technical insights and future directions in improving scalability along with a balance between both security and privacy. Section 5 discusses the available work related to blockchain scalabiltiy in comparison to this SLR. In the successive section, a brief discussion about blockchain technology is given. Research questions are identified to develop an understanding of data science approaches which will be a good direction for research community.

3 Blockchain overview

The term “blockchain” refers to a type of distributed-ledger-technology (DLT), which is a shared ledger with a continuously expanding list of entries that are kept and maintained in a “giant computer database”. This database is made up of several geographically unrestricted interconnected devices (phones, computers, or embedded systems) [31]. It contains an internal trust mechanism that is cryptographically enabled; thus network participants do not need to have mutual trust [32]. A block is the name given to each entry in this ledger, which is made up of messages and transactions and is connected and timestamped using cryptographic hashes [33] and network peers.

With the release of the Bitcoin white paper by Satoshi Nakamoto, the blockchain, a merger of two ancient technologies (cryptography and peer-to-peer communication) [2], gained popularity. Blockchain technology has been widely used as a result of the bitcoin blockchain, which was initially conceptualised in reaction to the global financial crisis of 2008 [3]. Speed, lower cost, security, fewer errors, fault tolerance, and the elimination of a central point of authority, attack, or failure are frequently highlighted by blockchain proponents as advantages over the current, largely centralized systems of operation in a number of industries outside of finance [3]. Other difficulties that restrict its scalability, meanwhile, have also emerged [30]. In Fig. 2, the line graph demonstrates the trend in the popularity of the term “blockchain” based on Google searches from 2010 to 2023 and in Fig. 3 we can observe how the popularity of the term “blockchain” has changed over time in different regions of the world.Understanding the trends in Google searches can provide insights into the overall public interest and awareness of blockchain technology during this time frame.

A special blockchain-based software platform called Ethereum makes it possible to create and operate DApps and smart contracts [5].

Ethereum, a special blockchain-based software platform, offers a complete programming language for the

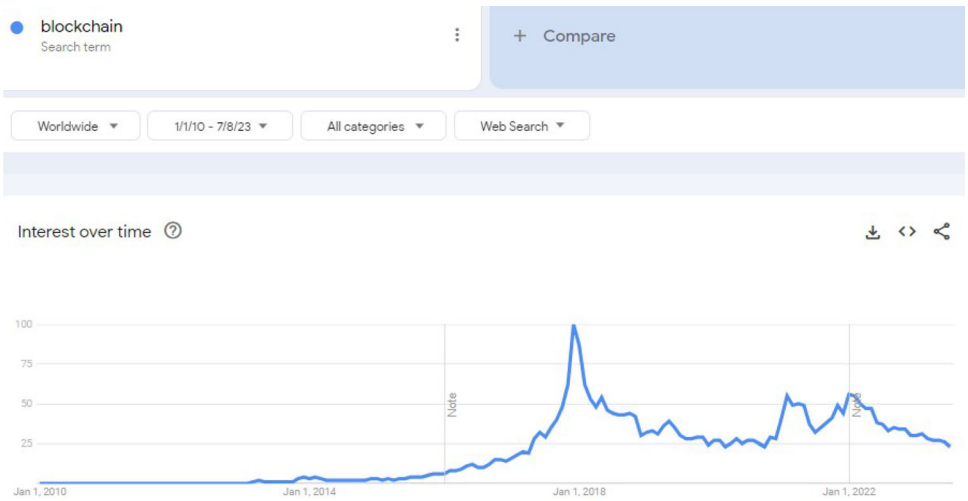


Fig. 2 Blockchain term Google trends 2010–2023

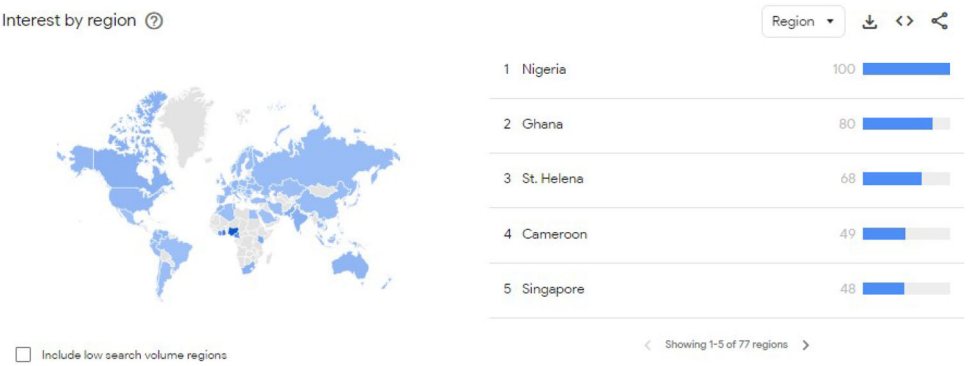


Fig. 3 "Blockchain" term Google trends 2010–2023

construction of smart contracts that enables writing programs and running them on the blockchain. Smart contracts allow developers to create complex applications that can automate tasks, verify the authenticity of data, and enforce the rules of the system. This makes it possible to create and operate DApps. Such platform serves as the foundation for the associated virtual currency known as Ether [33, 34, 39, 45, 47, 62, 33–34].

Since the Ethereum network is built around blocks, it is important to comprehend fundamentals of a block, to guarantee immutability and security, as it stores the blockchain network's basic data. In the next section, the components of a typical block is briefly explained.

3.1 Components of a block

Figure 4 depicts the components of a typical block on open blockchain networks. The Merkel tree binary hash mechanism is used by Bitcoin and Ethereum [10]. On these blockchains, a block is a hash value of data that must be documented. A header and the message body are the two

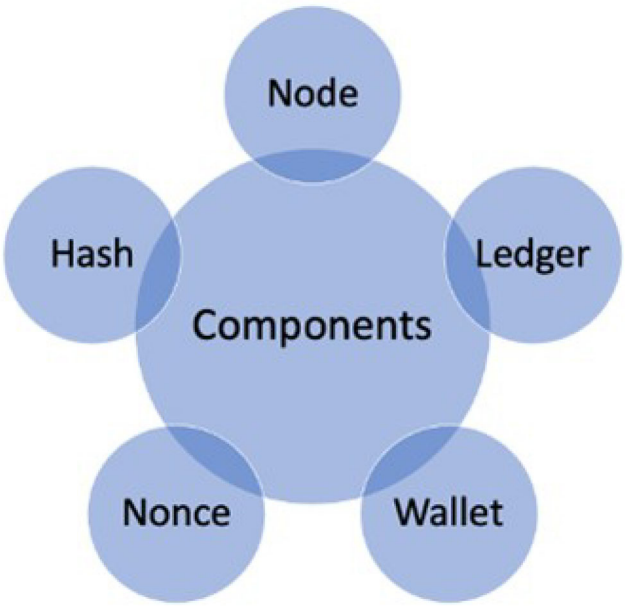


Fig. 4 Components of a typical block

main components of this data. The block version, Merkel root hash, parent hash, timestamp, nonce, and difficulty level are frequently included in the header. A list of transactions bundled together as the specific blockchain permits is found in the block body. Blockchain as a block contains numerous components for forming a network which includes below:

3.1.1 Node

In a network of transactions, there should be a component which maintains all transactions, and this is called Node. Node is primarily of two types i.e. full and partial. Full node contains a whole copy of all transactions being done under a network and has the capability of accepting, rejecting, and validating those transactions. On the other hand, a partial node just maintains the partial information of the transaction being done i.e., hash value (defined in Sect. 3.1.5).

3.1.2 Ledger

As the name implies, this is a digital database of information for the transactions happening around the blockchain network. It can be opened to the world, distributed, or decentralized in nature. The base permissions for Ledger transactions are read/write. This revolves around the action of transactions being done by a network of nodes.

3.1.3 Wallet

It is a digital tool for users to store, manage and interact with their digital assets on a blockchain network securely. Security is achieved by a private key which enables users to perform transactions and to maintain the integrity of their digital assets on the blockchain.

3.1.4 Nonce

A nonce stand for "number only used once," which is a 32-bit number added to a hashed or encrypted block in a blockchain, which makes transactions more secure.

3.1.5 Hash

Another important component of the blockchain network is Hash. Hash methods define the core security of the network and help to refine the transaction more deeply. On a blockchain network, a token, sometimes known as cryptocurrency, is occasionally needed for a transaction. These tokens can be used to reward network users for contributing the power, connection, and computing resources needed for network functioning. This cryptocurrency is referred to as *Bitcoin* on the Bitcoin blockchain and as *ether*, valued in gas, on the Ethereum blockchain. These cryptocurrencies frequently have values that are equal to those of fiat (traditional money, e.g., US dollars). Stable coin is a term used to describe cryptocurrencies that are based on local fiat money.

3.2 Key characteristics of blockchain

3.2.1 Decentralization

The necessity for each transaction to be verified by the central trusted agency (such as the central bank) in traditional centralized transaction systems [2] leads to cost and performance bottlenecks at the central servers. On the other hand, a transaction on the blockchain network may be carried out between any two peers (P2P) [6, 9, 30, 36] without the need for central agency authentication. Blockchain can help to alleviate performance constraints [8] at the central server and drastically lower server expenses (including development and operating costs).

3.2.2 Persistency

It is very hard to tamper with the network's transactions since each one must be verified and recorded in blocks [10] that are dispersed throughout the whole network. Each broadcasted block would also undergo transaction verification and validation by other nodes. Therefore, any fabrication can be immediately found.

3.2.3 Anonymity

With a created address, any user may communicate with the blockchain network. A user might also create many addresses to protect their identity. There is no longer a single entity in charge of protecting user privacy. With the help of this method, the transactions recorded on the blockchain [30, 31, 48] are kept somewhat private. Be aware that owing to an inherent restriction, blockchain cannot ensure full privacy protection.

3.2.4 Immutable

Once data is recorded on a blockchain, it cannot be altered or deleted. This makes it an ideal technology for tracking and verifying the authenticity of digital assets. Once a block is added to a blockchain, the data contained within it cannot be altered or deleted [42], making it a secure and reliable way to store and transfer information. This is achieved through the use of complex cryptographic techniques [5, 42, 43] and decentralized networks, which make it difficult for any one person or group to alter the data on

the blockchain. Additionally, Immutability can also be considered as the security feature of blockchain technology [30, 32, 33].

3.2.5 Transparency

Blockchain transparency refers to the ability of anyone to view and verify the transactions on a blockchain network. Blockchain technology is transparent, where all transactions are visible to all users on the network. This allows for greater accountability and trust among users [33]. This is achieved through the use of a distributed ledger (see Sect. 3.1.2), which is publicly accessible and contains a record of all transactions made on the network. Because of this, anyone can view the details of a transaction, including the sender, receiver, and amount involved, without the need for a central authority or intermediary. This transparency also allows for increased security and for easy detection of fraudulent activity. This transparency feature of blockchain enables its application to other systems like supply chain, government transaction, and voting systems [48]. Transparency [29] is one of the key advantages that blockchain technology offers over traditional systems, which may be opaque and relatively hard to audit.

3.2.6 Secure

Blockchain technology is highly secure, thanks to its use of cryptographic algorithms and consensus mechanisms [5, 42, 43, 88]. Transactions are verified and recorded on multiple copies of the blockchain, making it difficult for hackers to alter or corrupt the data. Blockchain technology is considered to be highly secure due to its decentralized and distributed nature. In a traditional centralized system, data is stored on a single server or a small number of servers, making it vulnerable to hacking and other forms of cyber attacks [12]. Blockchain technology, in contrast, uses a network of computers, or nodes, to store and verify transactions, making it much more difficult for hackers to compromise the system. One of the key security features of blockchain technology is the use of cryptography [5]. Each block in a blockchain contains a unique digital signature, or hash, which is created using complex mathematical algorithms. This hash is linked to the previous block in the chain, creating a chain of blocks that is extremely difficult to tamper with. In order to alter the data in a block, a hacker would have to change the hash of that block as well as all of the subsequent blocks in the chain, which is virtually impossible given the large number of nodes and computational power required.

It is important to note that no technology is 100% secure, but blockchain is considered to be relatively more secure technology, with levels of authentications, to store and transfer information and assets.

3.2.7 Smart contracts

Blockchain smart contracts are self-executing contracts with the terms of the agreement written directly into lines of code, stored and replicated on a blockchain network [5]. They can automate tasks, provide trust and transparency and reduce intermediaries. Smart contracts also provide a level of trust and transparency [37], as all parties can see the terms of the contract and the execution of the contract can be verified on the blockchain, which helps improve security and reduce the risk of fraud. However, they also have their own limitations like the need of correct code writing and the ability to change the smart contract itself.

3.2.8 Distributed

Blockchain technology is distributed, meaning that copies of the blockchain are stored on multiple devices across the network, making it resilient to single point of failure. Blockchain's distributed ledger technology allows multiple parties to record and validate transactions in a secure and transparent manner [29]. Instead of relying on a central authority to keep track of transactions, a blockchain network uses a decentralized system of computers, or "Nodes," to validate and record transactions [54] on a shared digital ledger. Each block in the chain contains a set of recent transactions, and once a block is added to the chain, its contents cannot be altered. This creates a tamper-proof record of all transactions on the network. This technology is used in many different applications, including cryptocurrency, supply chain management, and voting systems.

3.2.9 Interoperability

Interoperability in blockchain refers to the ability of different blockchain networks to communicate and exchange information with one another. This means that different blockchain networks can work together seamlessly, allowing users to transfer assets and information across different platforms. Blockchain can be integrated with other technologies such as IoT, AI, and cloud computing [37, 38] to create an ecosystem of interoperable systems. Interoperability is important as it enables greater flexibility and scalability in the use of blockchain technology. It also allows for the creation of cross-chain applications [48], where different blockchain networks can be connected to provide new and innovative services.

One way to achieve interoperability is through the use of 'Sidechains', which are separate blockchain networks that

are connected to the main blockchain. Another way is the use of 'Atomic Swap' which allows the exchange of assets between different blocks without the need for a trusted intermediary. Additionally, there are also several projects and protocols that are being developed to enable interoperability such as *Cosmos, Polkadot, CosmosSDK, Wanchain, and Aion* [31, 37, 47]. However, it is important to note that achieving interoperability is a complex task for which, there is ongoing research and development in this area, as well as standardization efforts to ensure that different blockchain networks can communicate in a consistent and secure manner.

3.3 Challenges of blockchain

As explained earlier, the security of a blockchain network and its scalability are often in a trade-off relationship. For example, Proof-of-Work (PoW) consensus mechanisms ensures strong security against attacks such as *Sybil attacks* and *double-spending*. However, PoW is computationally intensive, which limits the transactional throughput of the network, making it less scalable. On the other hand, Proof-of-Stake (PoS) consensus mechanisms are computationally less intensive, allowing higher transactional throughput, however, their security guarantees are weaker than PoW. The trade-off between security and scalability is a crucial consideration in designing and implementing blockchain networks. Major challenges of blockchain are security, scalability and privacy (see Fig. 5) and are brielfy explained below.

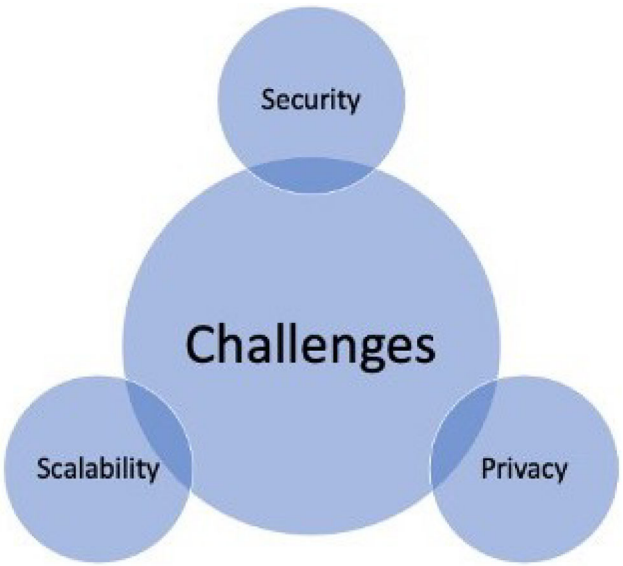


Fig. 5 Blockchain challenges

3.3.1 Security

Blockchain is much used technology in every field of interest, which poses an important security challenge to maintaining the hype. This includes consolidation of some real-common attacks which happen in the Blockchain network [87] such as *Liveness*, *Double spending*, *vulnerability*, *Private key security*, *Transaction privacy leakage*, *Selfish mining*, *Decentralized Autonomous Organization (DAO)* and *Border Gateway Protocol (BGP)*, *Hijacking*, *Balance* and *Sybil attacks* [3, 22, 89].

3.3.2 Privacy

Privacy becomes the primary concern of many users of Blockchain, when implemented into data-centric applications, that gives it the biggest challenge to overcome as per the use-case. As referred in [7], implementing Local Differential Privacy (LDP) strategies on the use-case model prevents the data privacy concern [7, 17, 33, 77, 90].

3.3.3 Scalability

As every technology face challenge with scalability and feasibility, blockchain is not an exception. Scalability in blockchain refers to the ability of a blockchain network to handle a large number of transactions and users without experiencing delays [3, 4] or performance issues. Since blockchain is distributed and decentralized in nature, scalability is another hurdle where researchers are working continuously to provide different solutions. Some of them are illustrated in [91, 92], where authors highlight the solutions by having roll-ups for network scalability in the chain of blockchain and a dynamic load-balancing technique for balancing the scalability of the network in Ethereum.

As more people and businesses adopt blockchain technology, the demand for faster and more efficient transaction processing increases. However, the current scalability of most blockchain networks is limited, with some networks, such as Bitcoin and Ethereum, only able to process a few transactions per second [81]. This is a major obstacle to the widespread adoption of blockchain technology.

There are several ways to improve the scalability of blockchain networks. One approach is to increase the block size, which is the amount of data that can be stored in each block on the blockchain. This allows for more transactions to be processed in each block, increasing the overall capacity of the network. However, increasing the block size can also lead to centralization, as it becomes more difficult for small and individual nodes to validate and process large blocks. Another approach is to use off-chain transactions [52, 53], which refers to moving transactions

off the main blockchain and on to a separate network. This can be done through the use of payment channels, such as the Lightning Network for Bitcoin, or state channels (e.g., *Raiden Network*) for Ethereum. These off-chain transactions are settled on the main blockchain periodically, allowing for a large number of transactions to be processed without congesting the main network.

Scalability solutions are still in the testing and implementation phase and it remains to be seen which approach will be the most effective in the long term. Additionally, scalability solutions often involve trade-offs and balance between decentralization and security [20, 28, 41, 47, 61, 93, 94], which also needs to be considered. Existing solutions have been discussed more in detail, in the later sections.

Summarizing, scalability is a critical issue that needs to be addressed for blockchain technology to reach its full potential. While there are various solutions proposed, it is important to strike a balance between scalability, security, and decentralization. As blockchain technology is still relatively new, it is likely that better scalability solutions will be developed in the future.

4 Data science

Data science is a multidisciplinary field that combines aspects of computer science, statistics, and domain expertise to extract insights and knowledge from data. The conventional disciplines are undergoing profound transformations due to the recognization of the significant challenges, possibilities, and substantial benefits of *Big Data*. Without charge, one can print or copy all or a portion of the work for their personal or educational use, provided that copies are made with the appropriate notice and are not manufactured or disseminated for profit or commercial gain. It is required to respect any copyrights for parts of the work that belong to somebody other than the author(s). Nonethless, Credit-assisted abstraction is acceptable. Moreover, scientific and technical sectors that revolve around data are also changing [71]. It is reshaping conventional data engineering fields including management, business, and social science.

Data is a driving force behind this reshaping and paradigm shift, but so are all the other things that may be made, changed, or improved by comprehending, utilising, and studying data. A fresh discussion concerning the so-called "fourth science paradigm," which unites experiment, theory, and computing (equivalent to "empirical" or "experimental," "theoretical," and "computational" science), has been sparked by the aforementioned trend and its potential. The future of science, technology, the economy, and potentially everything in our society now and tomorrow is driven by or even determined by data [74], which is seen as the new Intel processor, the new oil, and a strategic asset [95].

In Fig. 6, the line graph illustrates the trend in the popularity of the term "Data Science" based on Google searches from 2010 to 2023. Further, in Fig. 7, the visualization illustrates the variation in search interest for "Data Science" across different regions from 2010 to 2023. The map highlights the regions or countries where the search term "Data Science" has become popular. By analyzing this data, researchers and practitioners can identify geographic areas where Data Science has gained significant attention and where it may be less prevalent, facilitating a deeper understanding of the regional impact and adoption of Data Science as a discipline.

4.1 Branches in data science

There are several branches of data science, each with their own specific focus and set of tools and techniques. These branches include Data Mining, Machine Learning, Natural Language Processing (NLP), Computer Vision, Predictive Modeling, Big Data, Data Visualization and Data Engineering. We have only introduced the relevant data science branches which can be the solution to the scalability of blockchain and are mentioned below:

4.1.1 Machine learning

This branch of data science involves the use of algorithms and models to enable systems to learn from data and generate responses/predictions accordingly. It includes a wide range of techniques such as supervised learning [44], unsupervised learning, and reinforcement learning [52].

4.1.2 Predictive modeling

This branch of data science involves the use of statistical models and machine learning algorithms to make predictions about future events. It includes techniques such as linear regression, logistic regression, and decision trees [96–98].

4.1.3 Big data

Big data refers to the large and complex datasets that are generated by various sources such as social media, IoT devices, and e-commerce platforms. Big data as sub field of data science, deals with the management, analysis, and visualisation of extremely large and complex data sets. It requires powerful and specialized tools to manage, analyze, and visualize the data effectively. It includes tools such as Hadoop, Spark, Kafka [39] and NoSQL databases [42].

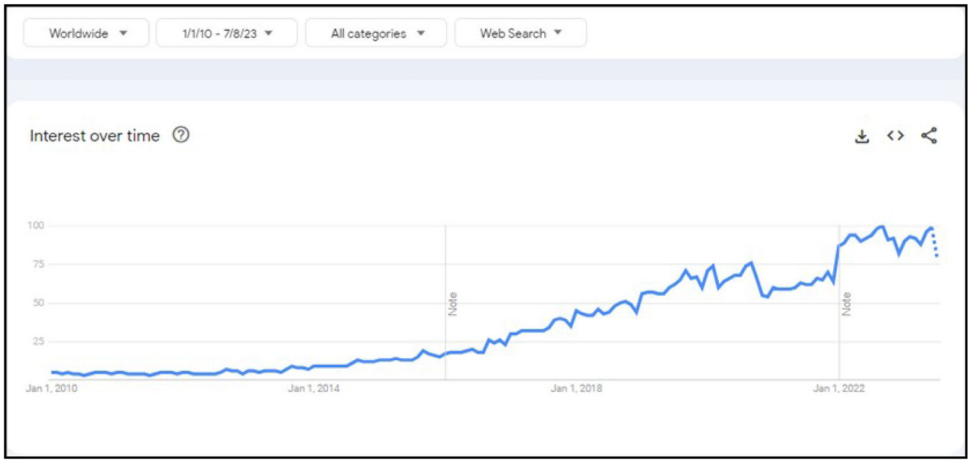


Fig. 6 Data science term Google trend from 2010–2023

Fig. 7 Data science term Google trends interest by region 2010–2023

4.2 Blockchain with big data tools

Blockchain and big data are two distinct technologies, but they can be integrated to create powerful solutions for various use cases. Blockchain can be used to encrypt and secure big data, making it more difficult for unauthorized parties to access or manipulate the data. Additionally, blockchain's decentralized nature can provide an extra layer of security by distributing the data across multiple nodes.

Big data tools such as *Hadoop* and *Spark* can be used to store and process data on a distributed blockchain network. This allows for data to be stored in a decentralized manner, which can improve data security and availability [99, 100]. Additionally, real time data analytics and streaming big data platforms, such as *Apache Kafka* and *Storm*, can be integrated with blockchain to enable real time data analytics. This can be useful for applications such as real time fraud detection and supply chain management. Some of the popular big data tools are briefed in following sub-sections.

4.2.1 Hadoop

Hadoop is an open-source framework for distributed storage and processing of large datasets. It consists of two main components; (1) *the Hadoop Distributed File System* (HDFS), which stores data across a cluster of commodity servers, and (2) the *MapReduce* [74] programming model, which processes the data in parallel.

4.2.2 Spark

Spark [49] is an open-source big data processing framework that is built on top of Hadoop. It is designed to be faster and more efficient than MapReduce, and it provides a wide range of libraries for data processing, machine learning, and graph processing. It can be integrated with blockchain to perform complex data processing tasks and to improve the scalability of the network. A research paper "Scalable analysis of the bitcoin blockchain using spark" analyze blockchain data by Rubin [85].

4.2.3 Kafka

Kafka [101] is an open-source distributed streaming platform which can be used to handle real time data feeds. It allows users to publish and subscribe to streams of data, and it can be integrated with blockchain to enable real time data analytics. A research paper " Mystiko—Blockchain Meets Big Data, Bandara [102] and "Rahasak—Scalable blockchain architecture for enterprise applications" by again Bandara [103] discusses how Kafka can be used to manage the data in a blockchain network, in order to improve its scalability and performance. It also allows users to publish and subscribe to streams of data, and can be integrated with other big data tools such as Storm and Spark.

It's worth noting that while these big data tools can be used to improve the scalability of blockchain networks, they are not a magic solution, and more research and development is needed to fully realize their potential. The choice of tool depends on the specific use case and the requirements of the project. Additionally, some companies offer cloud-based big data platforms, such as Amazon's *EMR*, Google's *Dataproc* and Microsoft's *HDInsight*, which allows users to easily spin up and manage big data clusters on cloud platforms.

Furthermore, the integration of blockchain with big data tools is still an emerging field and more research is needed to explore the best practices and the trade offs of these technologies.

5 Related work

There are claims that blockchain technology will upend the tech industry. However, due to previously mentioned scaling challenges, it has not yet succeeded in achieving this result. Numerous researchers have attempted to solve the scalability problem by putting forth various solutions [9, 32] yet the issue still exists. The challenge of scaling a blockchain and the evaluation of the offered solutions with various implementations have been the subject of several evaluations and surveys. Recently published surveys [59] evaluated the blockchain's scalability challenges in the healthcare sector and its implementation of healthcare systems, and authors provided workable solutions utilizing complementary data science and machine learning approaches. Major research papers for this SLR are categorized in Table 1 which provides a comprehensive overview of the baseline papers related to blockchain scalability. These papers serve as foundational works in the field, contributing to our understanding of the challenges and solutions pertaining to scalability in blockchain technology. The mentioned Table includes key information about each baseline paper, such as the title, author (1st) and publication year. By reviewing these baseline papers, individuals can gain valuable insights into the historical development of scalability concepts and the evolution of solutions proposed to address scalability challenges in blockchain networks. A brief explanation of each paper, listed in Table 1 are discussed below.

The comprehensive article [65], by Junfeng Xie, examines the broader spectrum of scalability issues in blockchain systems. The article offers a thorough evaluation of various scaling techniques and their effects on various blockchain systems. The scalability problem is attempted to be discussed from the viewpoints of throughput, storage, and networking by the authors. Then, current enabling technologies for scalable blockchain systems are discussed, along with relevant research hurdles and potential future research areas. Ahmad Akmaluddin in [59], examines the scalability issues that occur in the setting of blockchain technology with a focus on healthcare applications. The author examines *16* solutions, which may be divided into two categories, blockchain redesign and storage optimisation. However, there are still certain restrictions, such as those related to block size, large data

Table 1 Blockchain scalability baseline papers

Sr. No.	Year	Title	1st Author
1	2019	A Survey on the scalability of blockchain systems [65]	Junfeng Xie
2	2020	Scalability challenges in healthcare blockchain [59]	Ahmad Akmaluddin
3	2020	Solutions to scalability of blockchain: a survey [35]	Qiheng Zhou
4	2020	Scaling blockchains: a comprehensive survey [9]	Abdelatif Hafid
5	2020	Zecale: Reconciling Privacy and Scalability on Ethereum [104]	Antoine Rondelet
6	2021	Scalability improvement and analysis of permissioned-blockchain [38]	Swathi P.
7	2021	Systematic Literature Review of Challenges in Blockchain Scalability [37]	Dodo Khan
8	2023	Blockchain Scalability: Solutions, Challenges and Future Possibilities [105]	Moumita Roy

volume, transactions, number of nodes, and protocol difficulties. In this study, *13* alternatives for redesigning the blockchain, including blockchain modelling, read mechanisms, write mechanisms, and bi-directional networks, are assembled. There are three methods for storage optimisation. Guangsheng [35] examines several strategies for resolving scalability problems with blockchains. The study organises and evaluates several potential scalability fixes, ranging from on-chain and off-chain scaling methods to improvements in consensus mechanisms. This document makes an effort to list and categorise the current blockchain scaling options. Additionally, compares several approaches and gives a few probable routes for resolving the blockchain's scalability issue. Abdelatif [9], discusses several facets of blockchain scalability by taking into account elements like consensus methods, sharding, and layer 2 solutions, it offers insights into both the problems and solutions associated with scaling blockchain networks. More precisely, writers examined the most popular sharding-based blockchain protocols and suggested a taxonomy based on committee creation and intra-committee consensus. It also provides a performance-based comparison study of the benefits and drawbacks of the available scaling options (throughput and latency). Rondelet [104] propose a mechanism designed to balance privacy and scalability on the Ethereum blockchain. Zecale, a general-purpose SNARK proof aggregator that employs recursive composition of SNARKs, is presented by the authors, who contend that such scaling solutions for privacy-preserving state transitions are essential to simulate "cash" on blockchain systems. The trade-offs between privacy protections and scalability improvements have been explored by the authors and discuss how Zecale solves the above two critical issues. In [38], authors work on permissioned blockchains to make it more scalable, the author covers strategies and investigations that help hyperledger's Apache Kafka and Apache Spark networks scale more easily. In permissioned blockchains, scalability issues are investigated as potential solutions and methods. The SLR [37] gives a methodical literature research to fully comprehend the difficulties associated with scalability in blockchain technology. Its findings demonstrate that the Internet of Things (IoT) would be the leading application of blockchain in sectors like energy, finance, resource management, healthcare, education, and agriculture. The paper categorises and analyses various scalability solutions discovered in the literature. The scalability problems with these programmes, however, prevent them from producing the required results. Furthermore, the two main subcategories of scalability solutions are onchain and offchain. Onchain options include Sagwit, block size increase, sharding, and consensus techniques. On the other side, offchain is a lighting network. Moumita Roy in [105] highlight the issues and potential solutions related to blockchain scalability are covered by Moumita Roy in this study. In order to offer insight on the changing environment of scalability research, the author may examine new trends, technological developments, and potential future directions for improving the scalability of blockchain networks. Authors examined a number of the already available blockchain solutions based on sharding and provided a performance-based comparative study in the form of advantages and disadvantages of the current systems. Table 2 gives a brief comparison of the studies performed in the baseline papers with this SLR.

The rise of grant investigations, proposed applications and methodologies, technical reviews, are signs of blockchain's enormous potential [106–108]. In the future, implementation of blockchain technology, the number of healthcare projects and frameworks for security, privacy, and other financial sectors will increase [7, 80, 83]. The blockchain-based application, however, is a relatively recent invention. The quantity of false information, doubts, and speculative ideas regarding the possible applications of blockchain in the healthcare industry are explained [6, 7, 6–7, 6–7]. Notably, there are potentials and obstacles associated with blockchain-based applications specifically in the healthcare area, that require investigation from a variety of angles. Many healthcare procedures may be disrupted by a distributed system that removes the middlemen. One such idea is to combine a blockchain-based system with data science to increase scalability and use classification models for prediction [55, 57, 89, 98, 108, 116]. This is further discussed in the Sect. 7.

6 Research methodology

This section provides the methodology followed for conducting this SLR. The steps include identification of research questions, and carrying out research by screening selected work for literature, keywords and list of source data base. The identified research questions are then discussed in the next section.

6.1 Execution of systematic review

The steps of this review are as follows:

1. Formulation of the research question
2. Execution of the research techniques or procedures
3. Screening of research papers
4. Keywording based on the abstract
5. Data extraction

Table 2 Comparison of existing work with proposed study

S. No.	Title	Addressed issues	Scalability	Data science techniques	Application
1	A Survey on the scalability of blockchain systems [65]	Scaling with throughput, storage and network latency	Yes	No	General
2	Scalability challenges in healthcare blockchain [59]	Blockchain redesign and storage optimization	Yes	No	Healthcare
3	Solutions to scalability of blockchain: a survey [35]	On-chain and off-chain scaling methods for improving consensus mechanism	Yes	No	General
4	Scaling blockchains: a comprehensive survey [9]	Consensuc methods, sharding and layer 2 solutions	Yes	No	General
5	Zecale: Reconciling Privacy and Scalability on Ethereum [104]	A general purpose proof agregator for balancing privacy and scalability on ethereum blockchain.	Yes	Yes	General
6	Scalability improvement and analysis of permissioned-blockchain [38]	Strategies for hyperledgers Apache-Kafka and Spark for scaling	Yes	No	General
7	Systematic Literature Review of Challenges in Blockchain scalability [37]	Demonstrated the feasibility of IoT with various fields like energy, finance, healthcare, education and agriculture.	Yes	Yes	General
8	Blockchain Scalability: Solutions, Challenges and Future Possibilities [105]	Explores the existing solutions on sharding, and provides a performance based study of current and new trends in blockchain scalability	Yes	No	General
9	Proposed SLR	On-chain and Off-chain solutions to balance Scaling, Privacy and Security, with Network throughput and latency. Combine Blockchain with Machine Learning and Data Science approaches for better scalability.	Yes	Yes	General

6.2 Identification of the research questions

Before beginning the research methods, it is crucial to comprehend the research questions. The following are the research questions (RQ) regarding blockchain's scalability challenges, emphasizing data science and machine learning approaches.

RQ1: What are the current blockchain scalability issues?
RQ2: What are the proposed blockchain-based solutions to the problems?
RQ3: How researchers addressed the scalability issue with data science approach?

6.3 Conducting research

Implementing a variety of strategies, such as a search method, inclusion and exclusion criteria of literature and repository sources, is necessary to get the most meaningful findings.

6.3.1 Search strategy

The key term-based strategy for obtaining the linked articles' keywords was originally used in the study procedure. *"Blockchain"*, *"data science"*, and *"scalability"*, among other specific terms, were searched on Google Scholar. The articles were then picked and downloaded in order of first publishment. The aforementioned keywords related articles, papers, and reports were found in a range of journals, conferences on various repositories, e.g., IEEE, Science Direct, and SCOPUS, etc.

6.3.2 Exclusion and inclusion criteria

This SLR is developed with six primary eligibility requirements. The related literature inclusion and exclusion criteria are designed to exclusively accept materials or publications is visually presented in Table 3. The major focus is on the elements that are making it difficult to use blockchain on a broad basis and how the issue has been addressed with Data Science techniques. This SLR also targets publications that make an effort to address the scalability problem. Instead of only noticing the trend or notion without taking into account the implementation of solution feasibility, the proposed solutions should be constructed or simulated as well.

6.3.3 Screening of relevant articles

A number of citations given by the database were used to filter the identify related publications. It was essential to cite the pertinent papers at least once.

Table 3 Criteria for inclusion and exclusion in this SLR

Criteria	Title
1	The study must be an original investigation rather than a review or survey.
2	The publications that (directly or indirectly) address the topic of blockchain scalability and emphasise the pertinent causes or considerations.
3	The publications that (directly or indirectly) address the topic of blockchain scalability using data science techniques and emphasise the pertinent causes or considerations.
4	On the basis of formal evidence, simulation, and implementation, the proposed solutions have been assessed.
5	Peer-reviewed journals and conference journals publish the articles.
7	Only English should be used in the papers.

6.3.4 Keywording based on the abstract

After the abstract part, every article typically presents its most important keywords. Around *20* keywords were then collected from pertinent publications. However, the terms *"blockchain"*, *"scalability"*, and *"data science"*, were pertinent to this review.

6.3.5 Information and data sources

After lengthy discussion among study team members and authors based on the literature studied, the pertinent data and information repository sources were identified. As already stated, the primary emphasis of this research is the basic problem of scalability in public blockchains. For a high quality SLR research, authors relied on computer science literature in scholarly publications and conference proceedings. The data sources looked up include both substantial and interdisciplinary databases and specialized computer sciences. The following repositories provided articles for this study. These sources contain the broadest coverage of high quality literature in this field.

- Scopus
- IEEE Xplore
- Science Direct
- Web of Science

Each database/repository needs its own set of search terms. Appendix A lists the whole search strings for this SLR for the mentioned repositories. The PRISMA activity regulations stipulate that predefined search techniques are required to avoid potential bias during the publication search shown in Fig. 8. As a consequence, our search procedures were designed to execute the required literature using the internal search results of the above mentioned publication sites.

After conducting a number of thorough test searches, the keywords utilized in the search strings were determined. Initially, many databases were given the designations

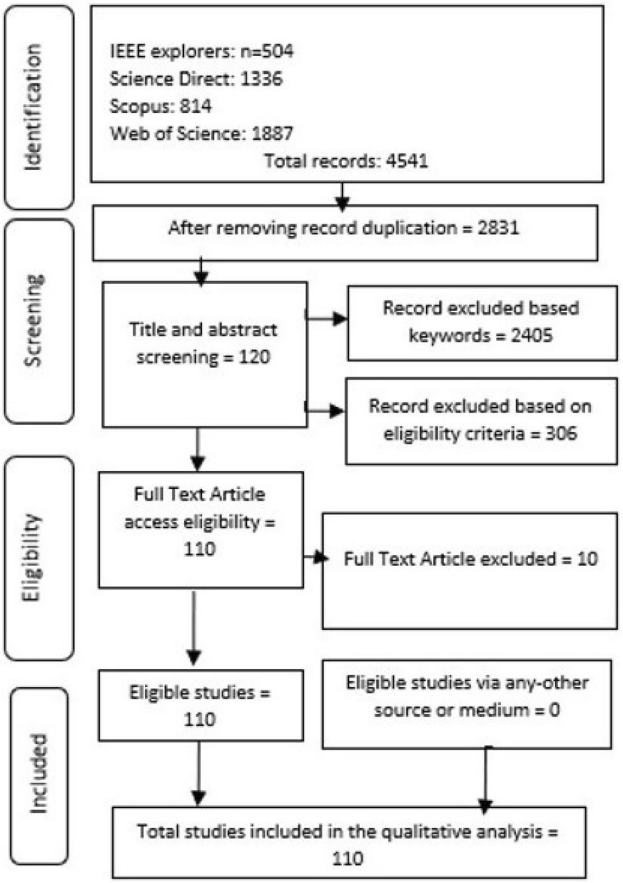


Fig. 8 Paper search and selection process (PRISMA Technique)

"Blockchain", *"Scalability"* and *"Data Science"*. Unfortunately, it showed how restricted in scope these keywords are. After experimenting with various keyword combinations, publications having clear technical synonyms for *"Blockchain"* and *"scalability"* were eventually found. This was done for articles between the years *2018* and *2022*. The *"scalability"* and *"data science"* search terms did not consistently produce a hit on every database. Using *"Blockchain"* and its technical synonyms, such as

"Ethereum", *"distributed ledger"*, *"apache Kafka"*, and *"apache spark"*, to conduct a search on IEEE Explore Library, it was discovered that the search produced publications that were closely related to the search criteria. These keywords, however, brought up a tonne of unrelated articles from other databases, most of which were about economics and/or cryptocurrencies, which were then shortlisted. In conclusion, this SLR contains almost all the research work published over these years, for the mentioned keywords.

Figures 9 and 10 show information on the various publications, including but not limited to, conference papers, journal papers, workshops and symposiums, that are included in this SLR.

7 Discussion

Blockchain research was minimal until 2016, most likely because it took some time for blockchain to take off after the debut of bitcoin in 2008. In 2016, specifically for public blockchains, research on blockchain scalability began to start and was visible in the research field. In 2016, there were more published papers on the topic of scalability; three years later, in 2019, there were more than 60. As time went on, blockchain started to significantly disrupt an increasing number of applications. Therefore, it makes sense for the research community to begin tackling the much discussed scalability issue in public blockchains.

The examination of around *110* articles that were chosen and published between the years *2018* and *2022* is covered in this section. It sheds light on the scalability problems, research trend over the previous four years and the solutions which exist for public and private blockchains. This section also illustrates and addresses the identified research questions,as stated in the Sect. 6.

7.1 RQ1: what are the current blockchain scalability issues?

Understanding the numerous blockchain scalability issues in the sector is crucial to this research subject. To identify

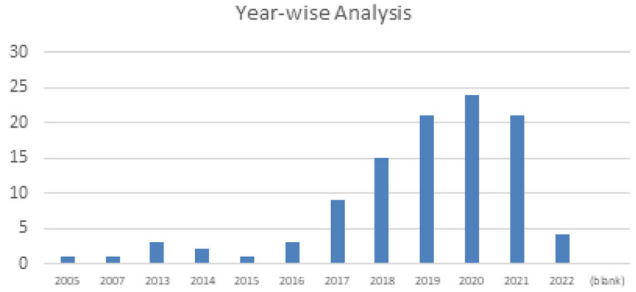


Fig. 9 Literature publications year-wise analysis

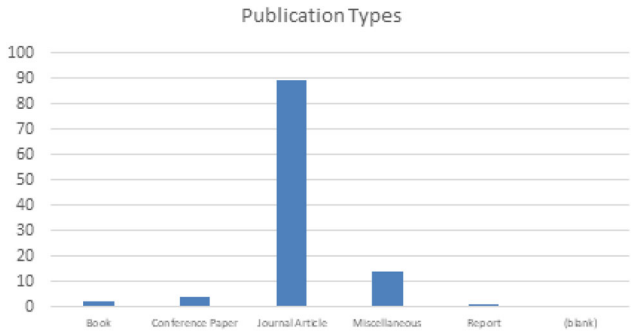


Fig. 10 Literature publications types analysis

the precise scalability issues that need to be solved, it is necessary to specifically study the pertinent publications from scientific databases. There have been several factors in the literature related to scalability from the literature i.e. Transaction Throughput, which can be defined as the number of transactions a blockchain can handle in one second [59] and network latency, which implies the amount of time required for the blockchain network to be ready for the next transaction [24, 94]. As per references, storage is also one of the factors, which refers to the storage capacity that a blockchain network can consume while the block size is the total capacity of the block to be utilized by the transactions, and blockchain network will probably reject the network if it exceeds the storage capacity. Table 4 presents a summary of scalability factors related to blockchain technology, as determined by the number of references cited in various research papers with each factor associated with a specific number of references.

There are several scalability issues that currently affect blockchain technology. Here are a few examples of scalability issues and their limitations:

7.1.1 Limited throughput

The limited throughput of most blockchain networks is a major scalability issue. This is due to the consensus mechanism used by most blockchains, which can only process a limited number of transactions per second. Research paper [117] discussed the limitations of the current consensus mechanism used by most blockchain networks and the scalability issues it creates.

7.1.2 Block size limitations

Most blockchain networks have a maximum block size, which limits the number of transactions that can be processed in each block. This can lead to network congestion and slow transaction processing times. Research paper by Dodo Khan [41] discusses the limitations of block size with the scalability issue.

Table 4 Scalability factors via No. of references

Sr. No.	Scalability factor	No of References
1	Throughput—This suggests the maximum number of transactions the protocol can process in a second.	23
2	Network Latency—This is relevant to the amount of time it takes to reach consensus after a transaction is launched.	15
3	Block size—This is the amount of space in a block that can be used by all transactions.	5
4	Storage—It speaks about the whole volume or capacity that a blockchain network may use.	2

7.1.3 High energy consumption

The proof-of-work consensus mechanism used by most blockchain networks is energy intensive, which can limit scalability and raise environmental concerns. Research paper [118] discussed the limitations of the proof-of-work consensus mechanism and the associated scalability issues.

7.1.4 Sidechains

Sidechains are separate blockchain networks that are connected to the main blockchain. This allows for interoperability and increased scalability. However, sidechains can also introduce security and complexity concerns. Research paper [94] discussed the limitations and challenges of sidechain-based solutions.

7.1.5 Limited privacy

Most blockchain networks provide a high degree of transparency and immutability, but this can also limit privacy and raise security concerns. [119] discussed the limitations of privacy in blockchain networks.

7.1.6 Complexity of smart contract

The complexity of smart contracts can lead to scalability issues, as the more complex the smart contract is, the more resources are required to execute it. Clack [120] discussed the limitations of smart contract complexity.

It's worth noting that these scalability issues are not mutually exclusive and can overlap with each other. Furthermore, Blockchain scalability is an active area of research and development, and several solutions are being proposed and developed to address these scalability issues. Figure 11 highlights the general existing scalability solutions.

7.2 RQ2: what are the proposed blockchain-based solution to the aforementioned problems?

Reviewing applicable papers reveals a variety of challenges that real world implementations encounter. Understanding and outlining the remedies in connection to the challenges is therefore crucial.

Scalability problems can be addressed through on-chain solutions by focusing on the block's internal components or off-chain solutions where the transactions can be processed outside the primary block (off-chain). These techniques, include Consensus Related Methods, Directed Acylic Graph (DAG) related methods, and some other potential solutions, like, Plasma, sharding, etc., proposed in the literature [59–65, 93], are briefly explained below.

7.2.1 Block-related methods

This method is On-chain solution, which is often large in block size. There are several important connections between the block size and the public blockchain scalability problem [101]. The huge block can surely support more transactions, which will undoubtedly increase throughput. This strategy is preferred by miners since it results in a bigger transaction fee while mining the block and allows for the inclusion of more transactions in a single block. With the Segwit mechanism [52], the block size is maintained while additional transactions are included in the block. In order to successfully create more room for subsequent transactions, this strategy seeks to separate the signature information first from the transaction and save it elsewhere [53]. If a transaction stays within a single shard, which is regarded as the biggest restriction, sharding is the only workable solution. Small shards, which are not scalable as they would produce a large number of shards, while big shards are also not scalable since the entire network implements the Byzantine consensus. Since, tiny fragments are more susceptible to security risks, each shard can withstand faults that are at most a third of its size. Additionally, if the number of malicious nodes stay constant, the shard's size would shrink, increasing the likelihood that it will fail.

7.2.2 Consensus related methods

For the purpose of enhancing scalability in public blockchains, these strategies have been put into practice and deployed in various applications. The key factor

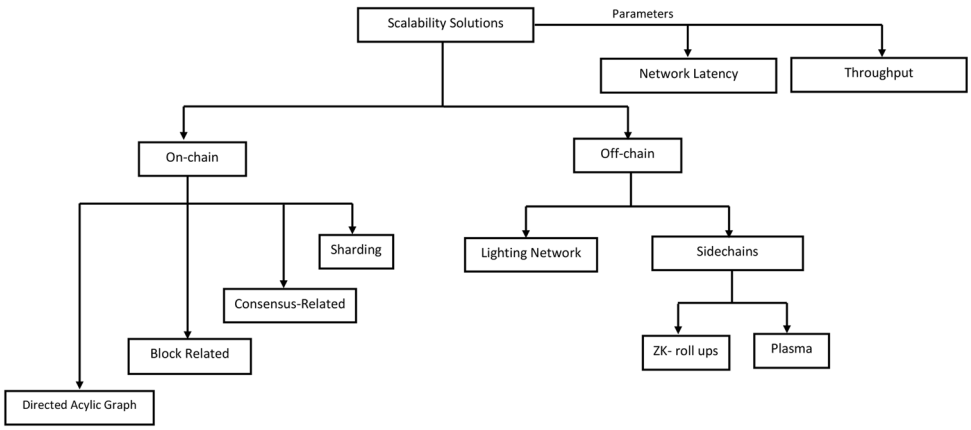


Fig. 11 Generalized blockchain scalability solutions

contributing to public blockchains' scalability problems is the consensus protocol's inefficiency. As a result, the research community has made an effort to find creative consensus solutions to the scaling problem. The scientific community has thus made an attempt to identify original consensus answers to the scaling challenge.

A smart contract serves as the structure for Plasma, a sidechain connected to the Ethereum network. With the primary chain, child chains are also produced. Periodically, the main chain posts the block headers of the Plasma sidechain for validation. Plasma's drawbacks include the necessity for the Ethereum network to review and validate sidechain blocks, which adds extra effort. The Ethereum blockchain must get all sidechain records in the event of a security breach (main chain). It is an off-chain approach that combines several transactions into a single light transaction that is stored on the main chain known as SNARK using Zero Knowledge (ZK) proof [104]. Sidechains conduct all executions, whereas the main Ethereum chain just records transaction data. ZK-roll restrictions include that it requires extensive computing work to produce. The primary drawbacks of ZK-rollup are centralization and apprehension about quantum computing being considered a temporary fix and does not offer smart contracts a simple migration path to layer 2.

7.2.3 Sharding

Data on a blockchain network is divided into smaller, more manageable units called "shards" by a process called "sharding." The processing and storage of each shard is then handled by a distinct node or group of nodes. Due to the ability to execute transactions in parallel, the network's total throughput is increased. Our findings show, Sharding, though, raises questions about the network's decentralization and security [47, 108].

7.2.4 Directed acylic graph related methods

DAG-based solutions [121] are a promising approach to address the scalability limitations of blockchain systems. In DAG-based distributed ledgers, transactions are represented as nodes in a directed acyclic graph (DAG), where each node references one or more previous nodes. This enables a more parallel and asynchronous processing of transactions compared to traditional blockchain systems, where transactions are added to a single linear chain. One of the most well known examples of DAG-based [122] distributed ledgers is IOTA Tangle. IOTA Tangle is designed specifically for the Internet of Things (IoT) and is optimized for high scalability, low resource consumption, and secure data exchange. Other DAG-based solutions, such as Hashgraph and Phantom, have also gained attention in recent years. However, DAG-based solutions still face some challenges, such as achieving consensus in a decentralized network and preventing attacks like double-spending. Nevertheless, the potential benefits of DAG-based solutions make them a promising area [123] for future research and development in blockchain scalability.

7.2.5 Off-chain transactions

Off-chain transactions let many transactions be performed without clogging up the primary blockchain network. Lightning Network may be established for an off-chain trade channel for the Bitcoin network, allowing two nodes to execute transactions quickly. The environment of the mining industry may alter, as transaction fees are declining or have completely gone. Another illustration that has been used in Ethereum is the Raiden network. It is referred to as the Ethereum based Lightning Network. With the exception of the transaction data, it operates using the same procedure and protocol as the Lightning network. Its state channels also transmit information about smart contracts.

Off-chain transactions' limitations and the fact that they are currently in the testing phase raise questions about their scalability and security. The restrictions and difficulties of off-chain transactions are mentioned in [15].

7.2.6 Plasma

Plasma is a scaling solution for blockchain networks that allows for off-chain computations and enables the creation of child-chains that are connected to the main blockchain network. This allows for increased scalability and security. However, Plasma also raises concerns about the complexity of the technology and the potential for centralization. Research paper [41] discussed the limitations and challenges of Plasma-based solutions.

It's worth noting that these solutions are still in the early stages of development, and more research and development is needed to fully understand their potential and limitations. Additionally, as the blockchain technology is constantly evolving, new solutions may arise and improve the current limitations. Although several applications have been put forth, not all of them have been successfully turned into a functional prototype. Real world implementations face several difficulties that may be found by reviewing pertinent publications. Therefore, it's critical to comprehend and map out the solutions to the problems. The research deficit would then be highlighted by these answers, which would also guide future research.

The categorized proposed solutions for blockchain scalability to our paper are mentioned in Table 5. It provides a taxonomy or classification of scalability in different layers of a blockchain system. The table likely includes multiple layers of the blockchain architecture, such as the application layer, consensus layer, and data layer, among others. For each layer, the table would categorize various scalability aspects, techniques, or solutions that have been studied or proposed in the literature.

7.3 RQ3: how researchers addressed the scalability issue with data science approach?

Researchers have addressed the scalability issue in data science by using a variety of techniques, such as distributed computing, parallel processing, and subsampling of data. Distributed computing allows for the data to be processed across multiple machines, reducing the memory and computational requirements on any single machine. Parallel processing allows for multiple algorithms to be run simultaneously on different subsets of the data. Subsampling of data involves randomly selecting a smaller subset of the data to train models on, which can significantly reduce the computational requirements. Additionally, researchers often use dimensionality reduction techniques to reduce the number of features in a dataset, which can also help with scalability.

There are several publications found addressing scalability of blockchain using data science approaches and several articles and work supported this approach. The research deficit is still to be noticed in it. In this SLR, the reviewed papers with data science techniques for blockchain have been reviewed. Major number of papers use framework and for a particular industry with Apache Kafka or Apache Spark, with major contributions towards the processing of data. Some papers have used machine learning models with blockchain frameworks to improve security and privacy and data management. To the best of my knowledge, only one paper directly addressing blockchain scalability i.e. Hyperledger scalability with Apache Spark and Apache Kafka were found and several industrial works recently began exploring Apache Kafka, Apache Spark, Machine Learning combinations with blockchain for better end frameworks, only due to the natural affinity of Apache Kafka and Blockchain. Apache Kafka platform also acts as the message broker in OmniPHR architecture, which uses its messaging and queuing features to exchange

Table 5 Taxonomy of the scalability in different layers

Blockchain layer	Categories of solutions	Existing solutions
Layer 2 (non-on-chain)	Payment channel	Light network, Raiden Network, Sprites
	Side chain	Pegged Sidechain, Plasma, Liquidity Network, ZK rollups
	Cross-chain	Cosmos, Polkadot
	Off-chain computation	Truebit, Arbitum
Layer 1 (on-chain)	Block data	Bitcoin—cash, Compact Block Relay
	Consensus	Bitcoin—NG, Algorand
	Sharding	Elastico, OmniLedger, Rapid chain, Monoxide
	DAG	Inclusive, Spectre,Phantom, Dagcoin

data between nodes. Apache Spark services are also applied to support scalable and responsive processing needs. LearnChain on Ethereum use comprehensive tests to highlight ita efficiency and effectiveness, however, it is a framework for security and privacy and not on scalability.

Reviewing through literature and to the best of authors' knowledge, there has been more research work with Apache Kafka and Apache Spark in the framework and as referred above only one paper by Dr. Swathi [50], addressed apache Kafka for the scalability of Hyper-ledger technology with efficient results. More work has been identified in 2022 and different research papers and industry work addressing the possibility of the immutable nature of Apache Kafka and Blockchain technology.

8 Results

The results of our analysis presented in Table 6 show that scalability remains a significant challenge in the blockchain field. Poor throughput, high transaction delays, and powerful energy consumption are the primary causes of low efficiency problems in popular blockchains such as Bitcoin and Ethereum. To address these challenges, data science techniques such as distributed computing, parallel processing, and subsampling of data can be used. These techniques can reduce the computational requirements and increase the processing speed of blockchain transactions. Furthermore, machine learning algorithms can optimize the consensus mechanisms of blockchain, improving the speed and efficiency of transaction processing. Our analysis also reveals that scalability in blockchain is not a single concept. There are several elements related to scalability, with transaction throughput and network latency receiving the most attention. The limitations of existing solutions such as sharding and off-chain scaling are also evident.

Overall, the combination of data science and blockchain technology offers a promising solution to address scalability challenges. Our findings suggest that further research is needed to identify and explore potential future directions for using data science techniques in blockchain, especially in the areas of distributed computing, parallel processing, and consensus mechanisms optimization.

9 Open issues and future direction

This section addresses the literature review challenges to be addressed and future research directions regarding blockchain, specifically related to scalability. While addressing the scalability of blockchain, security aspect needs to be taken care of, which remains a challenging task. Moreover, there exists an inverse relationship between scalability and privacy. This interdependent-ability brings limitations to blockchain scalability. Therefore, this SLR provides an overview of the available research regarding above mentioned scalability factors which need to be studied along with various data science approaches, and proves to be good research direction for the researchers.

Few major challenges, as recognized from the previous sections, are mentioned below:

- Limited resources to verify the scalability of current blockchain technology as only hyper-caliper has been used by the researchers, which is too challenging with ethereum implementation. Therefore, more adaptable solutions are required to validate the number of transactions per second (e.g., throughput and network latency) for better scalability.

Table 6 Results of analysis

Results of analysis	
Scalability	Significant challenge in the blockchain field
Causes of low efficiency	Poor throughput, high transaction delays, and high energy consumption in popular blockchains like Bitcoin and Ethereum
Techniques to address challenges	Distributed computing, parallel processing, and subsampling of data
Machine learning	Optimize consensus mechanisms to improve speed and efficiency
Scalability is not a single concept	Transaction throughput and network latency are the main concerns
Data science and blockchain	Offer promising solution to address scalability
Further research	Needed to explore potential future directions in distributed computing, parallel processing, and consensus mechanisms optimization

Table 7 Challenges and issues

Challenges and issues	Description
Limited resources available	More adaptable solutions are required to verify the number of transactions per second (e.g., throughput and network latency) for better scalability.
Focus on complex and energy-consuming solutions	Complex and huge energy consumption which is difficult for researchers with limited resources. Requires industrial involvement.
Data science and blockchain combination	The research is still in initial stage and the data science techniques can be explored to be combined with blockchain to tackle mentioned issues.

- Available scalability solutions are more complex and huge energy consumption which is difficult for researchers with limited resources, in most cases. Also, industrial involvement on blockchain scalability is needed, due to which resources can be available for implementations. This induistrial shift to blockchain can prove to be beneficial for industrial stake holders in long term due to distinguished features of blockchain (e.g., immutable smart contracts and enhanced security, as discussed in previous sections).
- Data Science approaches, like distributed cxomputing and parallel processing, can prove to be better for Blockchain to overcome various issues. Since, currently not much work is done in this domain. This was also the challenge for this SLR to sort relevant papers since many of available research was directed in different implementations and not with data science. Although, some blockchain techniques have been implemented in different industries like IoT, healthcare, and finance in terms of privacy and security and can be scaled as well, but the research is still in initial stage and the data science techniques can be explored to be combined with blockchain to tackle above mentioned issues.

The challenges mentioned above have a significant impact on research related to blockchain scalability and need to be addressed swiftly. The limited resources available to verify the scalability of current blockchain technology means that the number of transactions per second (throughput) and network latency cannot be adequately analyzed. This lack of analysis is a significant obstacle to the implementation and scaling of blockchain technology. Therefore, more adaptable solutions are required to verify blockchain parameters related to scalability. The complex solutions and high energy consumption models make it challenging for researchers to analyze and develop blockchain solutions. Although the combination of Data Science and Blockchain is promising, but existing work is inadequate in this domain. This lack of research makes it difficult to identify relevant papers and develop effective solutions. Table 7 shows above mentioned challeneges for easy understanding.

Based on the above discussion, it is evident that further research needs to be carried out to introduce, simple solutions, energy-efficient models, involvement of machine learning and data science techniques with blockchain and more solutions for transaction validation (i.e., network latency and throughput). Simple frameworks like Apache Kafka can be studied which has natural affinity and immutable logs which can improve scalability of blockchain and overcome various issues related to the three interdependent parameters of blockchain, (i.e., security, privacy and scalability).

10 Conclusion

In recent years, blockchain technology has gained significant attention due to its potential to revolutionize various industries. However, scalability remains a primary challenge hindering its widespread adoption. This review paper delves into the key challenges and solutions related to blockchain scalability emphasizing that scalability is a multifaceted issue that includes factors such as transaction throughput, storage, and network latency. The interdependence between these factors, which plays a critical role in blockchain scalability, has also been highlighted. The article acknowledges the limitations of the study and proposes future research directions, including the use of data science techniques such as predictive modeling and parallel processing to improve scalability. Furthermore, big data tools such as Apache Kafka and Apache Spark may be used to enhance scalability. The review also emphasizes the integration of blockchain with other emerging technologies such as data science and artificial intelligence to improve scalability. In conclusion, addressing blockchain scalability is crucial to realizing the technology's full potential in revolutionizing industries such as healthcare and finance. This article's exploration of the challenges and solutions related to scalability and the potential impact on various industries provides valuable insights for researchers, practitioners, and policymakers.

Appendix A: Section title of first appendix

An appendix contains supplementary information that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding of the research problem or it is information that is too cumbersome to be included in the body of the paper.

Acknowledgements The work is supported financially by the Ministry of Higher Education Malaysia via Fundamental Research Grant Scheme (FRGS/1/2019/ICT05/UM/01/1).

Author contributions Rao: Drafted the manuscript, including the introduction, methods, results, and discussion sections. Also, reviewed and revised the paper. M.K.: Provided supervision and guidance throughout the research process, and the content and paper were thoroughly discussed with her. MMH: Reviewed the paper and provided guidance in the results and discussion chapters. ZAM: Prepared the initial draft, worked on the journal's template, and created figures and tables.

Funding "This research was funded by the Ministry of Higher Education Malaysia via Fundamental Research Grant Scheme (FRGS/1/2019/ICT05/UM/01/1).

Data availability Enquiries about data availability should be directed to the authors.

Declarations

Confilict of interest The authors declare that they have no conflicts of interest. The results and conclusions of this research are solely those of the authors and do not represent the views or endorsement of any other individuals or organizations.

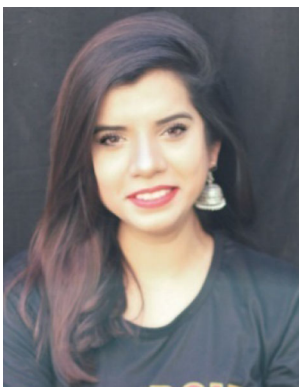
References

1. Wood, G., et al.: Ethereum: a secure decentralised generalised transaction ledger. Ethereum Proj. Yellow Pap. **151**(2014), 1–32 (2014)
2. Chauhan, A. et al.: Blockchain and scalability. In: 2018 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C), pp. 122–128. IEEE (2018)
3. Ochôa, I.S. et al.: Experimental analysis of the scalability of ethereum blockchain in a private network. In: Anais do II Workshop em Blockchain: Teoria, Tecnologia e Aplicações. SBC (2019)
4. Zmaznev, E.: Bitcoin and ethereum evolution. PhD thesis. Centria University of Applied Sciences (2017). https://www.theseus.fi/bitstream/handle/10024/141520/Thesis.pdf
5. Buterin, V. et al.: Ethereum white paper: a next generation smart contract & decentralized application platform. First version **53** (2014)
6. Shahbazi, Z., Byun, Y.-C.: Integration of blockchain, IoT and machine learning for multistage quality control and enhancing security in smart manufacturing. Sensors **21**(4), 1467 (2021)
7. Guangjun, W., et al.: Privacy-preserved electronic medical record exchanging and sharing: a blockchain-based smart healthcare system. IEEE J. Biomed. Health Inform. **26**(5), 1917–1927 (2021)
8. Sanka, A.I., Cheung, R.C.C.: A systematic review of blockchain scalability: issues, solutions, analysis and future research. J. Netw. Comput. Appl. **195**, 103232 (2021)
9. Hafid, A., Hafid, A.S., Samih, M.: Scaling blockchains: a comprehensive survey. IEEE Access **8**, 125244–125262 (2020)
10. Ferretti, S., D'Angelo, G.: On the ethereum blockchain structure: a complex networks theory perspective. Concurr. Comput.: Pract. Exp. **32**(12), e5493 (2020)
11. Wang, Z., Hu, Q.: Blockchain-based federated learning: a comprehensive survey. arXiv preprint arXiv:2110.02182 (2021)
12. Bez, M., Fornari, G., Vardanega, T.: The scalability challenge of ethereum: an initial quantitative analysis. In: 2019 IEEE International Conference on Service-Oriented System Engineering (SOSE), pp. 167–176. IEEE (2019)
13. Jabbar, A., Dani, S.: Investigating the link between transaction and computational costs in a blockchain environment. Int. J. Prod. Res. **58**(11), 3423–3436 (2020)
14. Rondelet, A.: Zecale: reconciling privacy and scalability on ethereum. arXiv preprint arXiv:2008.05958 (2020)
15. Ramanan, P., Nakayama, K.: Baffle: blockchain based aggregator free federated learning. In: IEEE International Conference on Blockchain (Blockchain), pp. 72–81. IEEE (2020)
16. Drungilas, V., et al.: Towards blockchain-based federated machine learning: smart contract for model inference. Appl. Sci. **11**(3), 1010 (2021)
17. Harris, J.D., Waggoner, B.: Decentralized and collaborative AI on blockchain. In: 2019 IEEE International Conference on Blockchain (Blockchain), pp. 368–375. IEEE (2019)
18. Awoke, T. et al.: Bitcoin price prediction and analysis using deep learning models. In: Communication Software and Networks: Proceedings of INDIA 2019, pp. 631–640. Springer (2020)
19. Liu, Y., et al.: Blockchain and machine learning for communications and networking systems. IEEE Commun. Surv. Tutor. **22**(2), 1392–1431 (2020)
20. Simpson, T. et al.: Fetch: Technical introduction. A decentralized digital world for the future economy (2018). https://fetch.ai
21. Van Otterlo, M.: A machine learning view on profiling. In: Privacy, Due Process and the Computational Turn-Philosophers of Law Meet Philosophers of Technology, pp. 41–64. Routledge, Abingdon (2013)
22. Hutchins, P.: Polygon Lightpaper. (2018). https://www.forbes.com/sites/forbestechcouncil/2018/10/02/creating-scalability-on-ethereum/#6eeefb575226
23. Harm, J., Obregon, J., Stubbendick, J.: Ethereum vs. bitcoin. www.economist.com (2016)
24. Kim, H., et al.: Blockchained on-device federated learning. IEEE Commun. Lett. **24**(6), 1279–1283 (2019)
25. Mohammed, A.H., Abdulateef, A.A., Abdulateef, I.A.: Hyperledger, Ethereum and blockchain technology: a short overview. In: 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), pp. 1–6. IEEE (2021)
26. Tun, M.T., Nyaung, D.E., Phyu, M.P.: Performance evaluation of intrusion detection streaming transactions using apache kafka and spark streaming. In: 2019 International Conference on Advanced Information Technologies (ICAIT), pp. 25–30. IEEE (2019)
27. Nie, JY.: Institute of Electrical and Electronics Engineers, and IEEE Computer Society. In: 2017 IEEE International Conference on Big Data: proceedings, pp. 11–14 (2017)
28. Jani, S.: An overview of ethereum & its comparison with bitcoin. Int. J. Sci. Eng. Res. **10**(8), 1–6 (2017)
29. Toyoda, K., et al.: Function-level bottleneck analysis of private proof-of authority ethereum blockchain. IEEE Access **8**, 141611–141621 (2020)

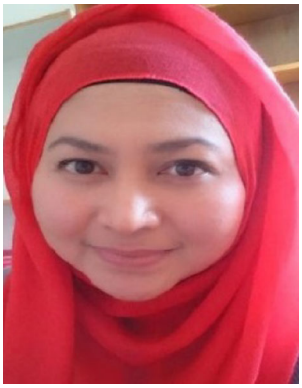
30. Zhang, L. et al.: Evaluation of ethereum end-to-end transaction latency. In: 2021 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS), pp. 1–5. IEEE (2021)
31. Gencer, A.E.: On scalability of blockchain technologies. PhD thesis. Cornell University (2017). https://search.proquest.com/docview/1964277559
32. Kanani, J. et al.: Polygon Lightpaper (2021). https://www.proquest.com/docview/1964277559
33. Croman, K. et al.: On Scaling Decentralized Blockchains Initiative for CryptoCurrencies and Contracts (IC3). http://fc16.ifca.ai/bitcoin/papers/CDE+16.pdf
34. Mahmood, Z., Jusas, V.: Implementation framework for a blockchainbased federated learning model for classification problems. Symmetry **13**(7), 1116 (2021)
35. Guangsheng, Y., et al.: Survey: sharding in blockchains. IEEE Access **8**, 14155–14181 (2020)
36. Chen, X. et al.: When machine learning meets blockchain: a decentralized, privacy-preserving and secure design. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 1178–1187. IEEE (2018)
37. Khalil, R. et al.: Commit-chains: secure, scalable off-chain payments. In: Cryptology ePrint Archive (2018). https://eprint.iacr.org/2018/642.pdf
38. Schäffer, M., Di Angelo, M., Salzer, G.: Performance and scalability of private Ethereum blockchains. In: Business Process Management: Blockchain and Central and Eastern Europe Forum: BPM 2019 Blockchain and CEE Forum, Vienna, Austria, September 1–6, 2019. Proceedings 17, pp. 103–118. Springer (2019)
39. Zhou, Q., et al.: Solutions to scalability of blockchain: a survey. IEEE Access **8**, 16440–16455 (2020)
40. Vujičić, D., Jagodić, D., Ranđić, S.: Blockchain technology, bitcoin, and Ethereum: a brief overview. In: 17th International Symposium Infoteh-Jahorina (infoteh), pp. 1–6. IEEE (2018)
41. Khan, D., Jung, L.T., Hashmani, M.A.: Systematic literature review of challenges in blockchain scalability. Appl. Sci. **11**(20), 9372 (2021)
42. Swathi, P., Venkatesan, M.: Scalability improvement and analysis of permissioned-blockchain. ICT Express **7**(3), 283–289 (2021)
43. Oliva, G.A., Hassan, A.E., Jiang, Z.M.: An exploratory study of smart contracts in the Ethereum blockchain platform. Empir. Softw. Eng. **25**, 1864–1904 (2020)
44. Benčić, F.M., Hrga, A., Žarko, I.P.: Aurora: a robust and trustless verification and synchronization algorithm for distributed ledgers. In: 2019 IEEE International Conference on Blockchain (Blockchain), pp. 332–338. IEEE (2019)
45. Abbas, K., et al.: A blockchain and machine learning-based drug supply chain management and recommendation system for smart pharmaceutical industry. Electronics **9**(5), 852 (2020)
46. Chen, S. et al.: A comparative testing on performance of blockchain and relational database: foundation for applying smart technology into current business systems. In: Distributed, Ambient and Pervasive Interactions: Understanding Humans: 6th International Conference, DAPI 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part I 6, pp. 21–34. Springer (2018)
47. Singh, S.K., Rathore, S., Park, J.H.: Blockiotintelligence: a blockchain-enabled intelligent IoT architecture with artificial intelligence. Future Gener. Comput. Syst. **110**, 721–743 (2020)
48. Frahat, R.T., Monowar, M.M., Buhari, S.M.: Secure and scalable trust management model for IoT P2P network. In: 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), pp. 1–6. IEEE (2019)
49. Safana, M.A., Arafa, Y., Ma, J.: Improving the performance of the proof-of-work consensus protocol using machine learning. In: 2020 Second International Conference on Blockchain Computing and Applications (BCCA), pp. 16–21. IEEE (2020)
50. Liu, X., Farahani, B., Firouzi, F.: Distributed ledger technology. Intelligent Internet of Things: From Device to Fog and Cloud, pp. 393–431 (2020)
51. Dobbelaere, P., Esmaili, K.S.: Kafka versus RabbitMQ: a comparative study of two industry reference publish/subscribe implementations: industry paper. In: Proceedings of the 11th ACM International Conference on Distributed and Event-Based Systems, pp. 227–238 (2017)
52. Borrero, J.D., Mariscal, J.: A case study of a digital data platform for the agricultural sector: a valuable decision support system for small farmers. Agriculture **12**(6), 767 (2022)
53. General Data Protection Regulation. General data protection regulation (GDPR). In: Intersoft Consulting. Accessed in October **24**(1) (2018)
54. Estupiñán, A.: Analysis of Modern Blockchain Networks Using Graph Databases. PhD thesis. Master's thesis, Technische Universitat Berlin (2020)
55. Choi, W., Hong, J.W.-K.: Performance evaluation of ethereum private and testnet networks using hyperledger caliper. In: 2021 22nd Asia-Pacific Network Operations and Management Symposium (APNOMS), pp. 325–329. IEEE (2021)
56. Dabbagh, M. et al.: Performance analysis of blockchain platforms: empirical evaluation of hyperledger fabric and ethereum. In: 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), pp. 1–6. IEEE (2020)
57. Iqbal, R., et al.: An experimental study of classification algorithms for crime prediction. Indian J. Sci. Technol. **6**(3), 4219–4225 (2013)
58. Venkatesan, N.J., et al.: Analysis of real-time data with spark streaming. J. Adv. Technol. Eng. Res. **3**(4), 108–116 (2017)
59. Mazlan, A.A., et al.: Scalability challenges in healthcare blockchain system—a systematic review. IEEE Access **8**, 23663–23673 (2020)
60. Schäffer, M., Di Angelo, M., Salzer, G.: Performance and scalability of private Ethereum blockchains. In: Business Process Management: Blockchain and Central and Eastern Europe Forum: BPM 2019 Blockchain and CEE Forum, Vienna, Austria, September 1–6, 2019, Proceedings 17, pp. 103–118. Springer (2019)
61. Chris, D.: Introducing Ethereum and Solidity Foundations of Cryptocurrency and Blockchain Programming for Beginners. Apress, New York (2017). https://doi.org/10.1007/978-1-4842-2535-6
62. Lewis, A.: Blockchain explained. In: Blockchain Technol. (2015). http://www.blockchaintechnologies.com/blockchain-definition
63. Ng, W.Y., et al.: Blockchain applications in health care for COVID-19 and beyond: a systematic review. Lancet Digit. Health **3**(12), e819–e829 (2021)
64. Chukwu, E., Garg, L.: A systematic review of blockchain in healthcare: frameworks, prototypes, and implementations. IEEE Access **8**, 21196–21214 (2020)
65. Xie, J., et al.: A survey on the scalability of blockchain systems. IEEE Netw. **33**(5), 166–173 (2019)
66. Rouhani, S., Deters, R.: Performance analysis of ethereum transactions in private blockchain. In: 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 70–74. IEEE (2017)
67. Memon, R.A., Li, J.P., Ahmed, J.: Simulation model for blockchain systems using queuing theory. Electronics **8**(2), 234 (2019)

68. Memon, R.A., et al.: Cloud-based vs. blockchain-based IoT: a comparative survey and way forward. Front. Inf. Technol. Electron. Eng. **21**(4), 563–586 (2020)
69. Memon, R.A., et al.: DualFog-IoT: additional fog layer for solving blockchain integration problem in Internet of Things. IEEE Access **7**, 169073–169093 (2019)
70. Memon, R.A. et al.: Modeling of blockchain based systems using queuing theory simulation. In: 2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 107–111. IEEE (2018)
71. Donawa, A., Orukari, I., Baker, C.E.: Scaling blockchains to support electronic health records for hospital systems. In: 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp. 0550–0556. IEEE (2019)
72. Gao, Z. et al.: Scalable blockchain based smart contract execution. In: 2017 IEEE 23Rd International Conference on Parallel and Distributed Systems (ICPADS), pp. 352–359. IEEE (2017)
73. Blanchard, P. et al.: Machine learning with adversaries: Byzantine tolerant gradient descent. Adv. Neural Inf. Process. Syst. **30** (2017)
74. Singla, K., Bose, J., Katariya, S.: Machine learning for secure device personalization using blockchain. In: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 67–73. IEEE (2018)
75. Mugunthan, V., Rahman, R., Kagal, L.: Blockflow: an accountable and privacy-preserving solution for federated learning. arXiv preprint arXiv:2007.03856 (2020)
76. Li, Y., et al.: A blockchain-based decentralized federated learning framework with committee consensus. IEEE Netw. **35**(1), 234–241 (2020)
77. Nagar, A.: Privacy-preserving blockchain based federated learning with differential data sharing. arXiv preprint arXiv:1912.04859 (2019)
78. Chen, P., et al.: Research on scalability of blockchain technology: problems and methods. J. Comput. Res. Dev. **55**(10), 2099–2110 (2018)
79. Bouoiyour, J., Selmi, R.: Ether: bitcoin's competitor or ally? arXiv preprint arXiv:1707.07977 (2017). http://arxiv.org/abs/1707.07977
80. Antwi, M.S., et al.: The case of HyperLedger Fabric as a blockchain solution for healthcare applications. Blockchain: Res. Appl. **2**(1), 100012 (2021)
81. Saudi Computer Society. In: 2nd International Conference on Computer Applications & Information Security (ICCAIS' 2019). Riyadh, Kingdom of Saudi Arabia (2019)
82. Gurusamy, V., Kannan, S., Nandhini, K.: The real time big data processing framework: advantages and limitations. Int. J. Comput. Sci. Eng. **5**(12), 305–312 (2017)
83. Roehrs, A., et al.: Analyzing the performance of a blockchain-based personal health record implementation. J. Biomed. Inform. **92**, 103140 (2019)
84. Omar, I.A., et al.: Supply chain inventory sharing using ethereum blockchain and smart contracts. IEEE Access **10**, 2345–2356 (2021)
85. Rubin, J.: Btcspark: scalable analysis of the bitcoin blockchain using spark. Dec **16**, 1–14 (2015)
86. Wang, K., et al.: Securing data with blockchain and AI. IEEE Access **7**, 77981–77989 (2019)
87. Singh, S., Hosen, A.S.M.S., Yoon, B.: Blockchain security attacks, challenges, and solutions for the future distributed iot network. IEEE Access **9**, 13938–13959 (2021)
88. Blockchain-based security management of IoT infrastructure
89. Zhang, Z., et al.: Recent advances in blockchain and artificial intelligence integration: feasibility analysis, research issues, applications, challenges, and future work. Secur. Commun. Netw. **2021**, 1–15 (2021)
90. Bao, X. et al.: Flchain: a blockchain for auditable federated learning with trust and incentive. In: 2019 5th International Conference on Big Data Computing and Communications (BIGCOM), pp. 151–159. IEEE (2019)
91. Thibault, L.T., Sarry, T., Hafid, A.S.: Blockchain scaling using rollups: a comprehensive survey. In: IEEE Access (2022)
92. Wang, Z., Cui, B., Hou, W.: A dynamic load balancing scheme based on network Sharding in private Ethereum blockchain. In: 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 362–367. IEEE (2022)
93. Dhulavvagol, P.M., Bhajantri, V.H., Totad, S.G.: Blockchain ethereum clients performance analysis considering E-voting application. Procedia Comput. Sci. **167**, 2506–2515 (2020)
94. Johnson, S., Robinson, P., Brainard, J.: Sidechains and interoperability. arXiv preprint arXiv:1903.04077 (2019)
95. Cao, L.: Data science: a comprehensive overview. ACM Comput. Surv. (CSUR) **50**(3), 1–42 (2017)
96. Yoo, Y.: The tables have turned: how can the information systems field contribute to technology and innovation management research? J. Assoc. Inf. Syst. **14**(5), 227 (2013)
97. Dinh, T.T.A. et al.: Blockbench: a framework for analyzing private blockchains. In: Proceedings of the 2017 ACM International Conference on Management of Data, pp. 1085–1100 (2017)
98. Sandner, P., Gross, J., Richter, R.: Convergence of blockchain, IoT, and AI. Front. Blockchain **3**, 522600 (2020)
99. Kurtulmus, A.B., Daniel, K.: Trustless machine learning contracts; evaluating and exchanging machine learning models on the ethereum blockchain. arXiv preprint arXiv:1802.10185 (2018)
100. Kim, H. et al.: On-device federated learning via blockchain and its latency analysis. arXiv preprint arXiv:1808.03949 (2018)
101. Thein, K.M.M.: Apache kafka: next generation distributed messaging system. Int. J. Sci. Eng. Technol. Res. **3**(47), 9478–9483 (2014)
102. Bandara, E. et al.: Mystiko—blockchain meets big data. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 3024–3032. IEEE (2018)
103. Bandara, E., et al.: Rahasak-scalable blockchain architecture for enterprise applications. J. Syst. Archit. **116**, 102061 (2021)
104. Rondelet, A.: Zecale: reconciling privacy and scalability on ethereum. arXiv preprint arXiv:2008.05958 (2020). http://arxiv.org/abs/2008.05958
105. Roy, M., Singh, M., Radhakrishnan, B.: Blockchain scalability: solutions, challenges and future possibilities. In: International Conference on Signal & Data Processing, pp. 133–149. Springer (2022)
106. Chan, W., Olmsted, A.: Ethereum transaction graph analysis. In: 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST), pp. 498–500. IEEE (2017)
107. Zheng, Z., et al.: Blockchain challenges and opportunities: a survey. Int. J. Web Grid Serv. **14**(4), 352–375 (2018)
108. Chen, F., et al.: Machine learning in/for blockchain: future and challenges. Can. J. Stat. **49**(4), 1364–1382 (2021)
109. Lee, H.-A., et al.: An architecture and management platform for blockchainbased personal health record exchange: development and usability study. J. Med. Internet Res. **22**(6), e16748 (2020)
110. Choi, Y., et al.: Development of a mobile personal health record application designed for emergency care in Korea; integrated information from multicenter electronic medical records. Appl. Sci. **10**(19), 6711 (2020)

111. Hussien, H.M., et al.: Blockchain technology in the healthcare industry: trends and opportunities. J. Ind. Inf. Integr. **22**, 100217 (2021)
112. Zhuang, Y., et al.: Generalizable layered blockchain architecture for health care applications: development, case studies, and evaluation. J. Med. Internet Res. **22**(7), e19029 (2020)
113. Roehrs, A., Da Costa, C.A., da Rosa Righi, R.: OmniPHR: a distributed architecture model to integrate personal health records. J. Biomed. Inform. **71**, 70–81 (2017)
114. Chang, R.-I., et al.: Blockchain for bounded-error-pruned content protection. ICT Express **7**(3), 295–299 (2021)
115. Balistri, E., et al.: BlockHealth: blockchain-based secure and peer-to-peer health information sharing with data protection and right to be forgotten. ICT Express **7**(3), 308–315 (2021)
116. Wang, Z. et al.: Kafka and its using in high-throughput and reliable message distribution. In: 2015 8th International Conference on Intelligent Networks and Intelligent Systems (ICINIS), pp. 117–120. IEEE (2015)
117. Eyal, I. et al.: Bitcoin-NG: a scalable blockchain protocol. In: 13th USENIX symposium on networked systems design and implementation (NSDI 16), pp. 45–59 (2016)
118. De Vries, A.: Bitcoin's growing energy problem. Joule **2**(5), 801–805 (2018)
119. Moser, M.: Anonymity of bitcoin transactions (2013)
120. Clack, C.D.: Smart contract templates: legal semantics and code validation. J. Digit. Bank. **2**(4), 338–352 (2018)
121. Fan, C. et al.: Towards a scalable DAG-based distributed ledger for smart communities. In: 2019 IEEE 5th World Forum on Internet of Things (WFIoT), pp. 177–182. IEEE (2019)
122. Gangwani, P., et al.: Securing environmental IoT data using masked authentication messaging protocol in a DAG-based blockchain: IOTA tangle. Future Internet **13**(12), 312 (2021)
123. Wang, Q., et al.: Sok: Dag-based blockchain systems. ACM Comput. Surv. **55**(12), 1–38 (2023)



Iqra Sadia Rao completed her Bachelors in Mechatronics Engineering, which included an exchange semester at Kansas State University in the USA. Recently, she completed her Masters in Computer Science from the prestigious University of Malaya in Kuala Lumpur, which is a top-ranked university according to QS World University Rankings. She developed a chatbot that raised awareness about mental health issues arising during the COVID-19 pandemic, which was acknowledged and appreciated by the US State Department's Morgan Ortagus in 2020. She has worked as a Data Professional for 7 years at International level working with multinational companies. She is currently working on a research project "An Improved Solution on sidechain manipulation for secure and scalable Ethereum's Blockchain by Data Science Techniques" which she is implementing in the Health Sector under the Ministry of Malaysia. She has published several conference papers, presented a paper in Italy recently, eTELEMED 2023 and journal papers focusing on her research interests include Blockchain, Artificial Intelligence and Data Science. Rao: Drafted the manuscript, including the introduction, methods, results, and discussion sections and later revised the paper.



M. L. Mat Kiah contributed to supervise this research paper and content and paper was thoroughly discussed with her. She received her PhD degree in Information Security from Royal Holloway, University of London, United Kingdom in 2007, and since then she is an active researcher at Faculty of Computer Science & Information Technology, UM in her Computer Science field particularly in Security. She was promoted to Professorship in 2015, and is an active member of IEEE as Senior Member, EC Council, Malaysian Society for Cryptology Research (MSCR) and Malaysia Board of Technologists (Ts.). Her main research interest will always be in the Security aspect of Computing and Technology fields with variation of applications in multi and/or trans disciplinary projects. This is evidenced by her publications and research projects in which she is/was the principal investigator (PI) as well as co-PIs. As a professional technologist (Ts.), keeping up with the current trend and demand of ever evolving Computing Technology field is crucial to ensure the quality and the impact of her research work. Current research interests include Cyber Security, Blockchain Technology, IoT and Health Information Exchange. M.K: Provided supervision and guidance throughout the research process, and the content and paper were thoroughly discussed with her.

M. Muzaffar Hameed is an Assistant Professor (Department of Computer Science) & Director IT at Bahauddin Zakariya University, Multan. He has finished his Ph.D from the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia in 2023. He gained a Bachelor in Computer Science (BSCS) from BZU Multan, Pakistan, and a Masters in Software Engineering from BTH, Karlskarona, Sweden. He has 12 years of experience related to software development, teaching and research. He has been awarded a foreign PhD scholarship by HEC, Pakistan. His field of research is machine learning, deep learning, digital forensics and information security. Hameed: Reviewed the paper and provided guidance in the results and discussion chapters.

Zain Anwer Memon recieved bachelor's and master's degrees in electronic engineering from Mehran University Jamshoro. He obtained dual Ph.D. degrees in Electronic and Communication Engineering from Politecnico di Torino, Italy, and Xi'an Jiaotong University, China, in 2013, 2016, and 2021, respectively. Currently, he serves as an Assistant Professor and Head of the Department of Electronics and Telecommunication Engineering at the University of Sindh, Jamshoro. Dr. Memon was also an exchange scholar at St. Cloud State University, USA (Fulbright program) and Frederick University, Cyprus (Erasmus Mundus scholarship). His research focuses on power grid networks, load-flow solutions, statistical assessment of next-gen power networks, surrogate modeling, blockchain networks, and data reduction techniques. He received the Best Student Paper Award at the 6th IEEE Global Electromagnetic Compatibility Conference (GEMCCON 2020) during his Ph.D.