The benefits and challenges of applying Blockchain technology into Big Data: A literature review

Thu Nguyen[1,2], Khoa Tan VO[1,2], Thu-Thuy Ta[1,2], Tu-Anh Nguyen-Hoang[1,2], Ngoc-Thanh Dinh[3]
[1]*University of Information Technology, Ho Chi Minh City, Vietnam*
[2]*Vietnam National University, Ho Chi Minh City, Vietnam*
[3]*Soongsil University, Seoul, South Korea*
{thunta,khoavt,thuyta,anhnht}@uit.edu.vn, thanhdcn@dcn.ssu.ac.kr

***Abstract*—Blockchain and big data are two technologies with great potential and great influence in the field of information technology. Big data research applied in many fields of society. However, big data has features as large size, temporal events, complex structure, and incompleteness. Therefore, big data exists many challenges that need to be researched such as data security, data integrity, anti-fraud, data quality, data management, data analysis, and data mining. Blockchain technology has the characteristics of distribution, immutability, transparency, and security. Therefore, integrating blockchain technology into big data is a promising solution to overcome these challenges. However, blockchain technology is not really mature yet. Researchers need to identify the problem and have a suitable approach for applying blockchain technology to big data. In this article, we survey and present a complete picture of the integrated base. At the same time, cloud services for big data, application range, and blockchain big data projects also presented. From there, the researchers were able to clearly identify the development challenges and future directions for this exciting topic.**

***Index Terms*—Blockchain, Big Data, Security.**

I. Introduction

Big data is one of the most prominent information technology topics today. Today, big data applied in many social fields. The big data market will grow rapidly in the coming years. However, big data has special characteristics and still has many problems when applied in reality. Big data is everything (information about every aspect of life), quantification (mostly information stored in digital form) and tracking (analyzed and updated continuously over time) [1]. Therefore, big data often has characteristics: large size generated by billions of users, events over time, complex and incomplete data structures. Researchers have conducted many studies on the characteristics, exploitation potential and future approaches of big data. The key challenges of big data are ensuring privacy, data security, data integrity, fraud, data analysis, data sharing, data quality, and data access.

Blockchain is one of the most prominent technologies today. This technology opens a new opportunity to improve the performance of other fields such as supply chain, finance, smart city, etc [2]. With the combination of cryptography, distributed computing, consensus algorithm and peer-to-peer network, blockchain technology is a distributed network with strong security. Data stored on the blockchain network has special characteristics: immutability, transparency, and security. Therefore, blockchain technology has the right characteristics to solve the problems of big data.

First, we conduct a survey and analyze the core challenges of big data. We then explore the challenges that remain when using cloud services for storage, retrieval, and analysis. Next are the challenges when applying big data in some prominent fields today. A general picture of the challenges of big data presented in Fig. 1.

In this article, we will analyze the basis of integrating blockchain technology and big data in section II. We then learn about the challenges of big data when deploying on a cloud service and present the current blockchain solution in section III. Section IV presents blockchain solutions that help overcome the challenges of big data when applied in different fields. At the same time, we also present some outstanding blockchain for big data projects today in section V. Section VI presents development challenges and future directions. Finally, we summarize the survey in section VII.

II. Integration facility

A. Ensuring data privacy and security

Problems:

- Large volume of data.
- Multiple data types.
- Data constantly updated from the multiple sources.
- Cloud-based data storage.
- Distributed processing of data.

Trends in cybersecurity breaches, vulnerabilities and third-party dependencies create challenges in terms of data privacy and security [3]. Traditional security solutions such as firewalls, number of hits and access rights, etc. are no longer effective because big data is beyond the control scope of organizations [4]. Big data collected from many sources and stored in many places, so it is not completely dependent on the systems and networks of organizations.

Solution: Blockchain network stores data decentralized but still ensures data privacy and security thanks to encryption mechanism and the right to monitor all transactions in a peer-to-peer manner.

B. Data integrity

Problems:

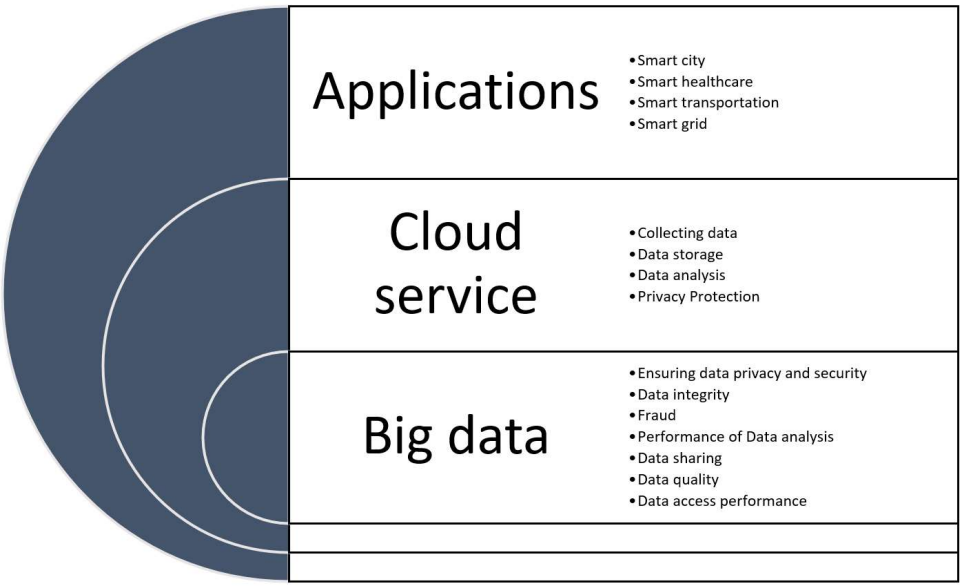

Fig. 1. Challenges of big data.

- Multiple processes involved in the data lifecycle [5].
- Data quality on cloud storage services [6].
- Data quality from IoT devices have highly dynamic and resource-constrained [5].

Data from multiple processes does not meet the constraints. Data from cloud storage services needs to be monitored to avoid information modification. Security vulnerabilities of IoT devices make data quality unreliable. Therefore, these factors affect the integrity of big data.

Solution: Blockchain technology provides a solution to store data with integrity and immutability in a distributed environment. Data stored on the blockchain network cannot be modified and tampered with. Therefore, the data ensures reliability and integrity.

C. Fraud

Problems:

- The process of fraud detection consists of fraud that has occurred in practice and fraud that is expected to occur. Traditional data mining techniques and current big data solutions almost based on transaction history [7].
- Real-time transaction monitoring is not possible [4].

Detecting financial fraud is difficult and challenging. Current data mining techniques and big data solutions cannot solve all these difficult problems.

Solution: The consensus algorithm and smart contract in blockchain technology make the process of accessing and processing transactions faster than traditional methods. Storing big data in the blockchain network makes it possible for organizations to monitor financial transactions in real time. Therefore, this technological solution can help improve the efficiency of financial fraud detection.

D. Performance of Data analysis

Problems:

- Taking a long time to process and prepare data [8].
- Poor quality data and errors make it difficult to process and filter data [8].
- In the financial sector, cross-border transactions are difficult to control and process [4].

Performance of data analysis is not high and real-time transaction monitoring is not possible.

Solution: Applying blockchain technology to big data analysis helps cross-border transactions can be processed in near real time [4]. The characteristics of blockchain technology could solve the problem of performance of data analysis if big data stored on the blockchain network. Blockchain technology uses consensus algorithms and smart contracts to make data retrieval easy and possible in real time.

E. Data sharing

Problems:

- Big data stores large amounts of sensitive user data [9].
- Enterprises share sensitive user data to reduce the cost of building personalized services and data services [9].

Sharing sensitive data on big data platforms securely is a big challenge.

Solution: Blockchain blocks linked, encrypted, and stored on a distributed ledger in the blockchain network, making data transmission between stakeholders convenient and minimizing many risks. All data recorded on a distributed ledger, avoiding duplicate data sharing and analysis.

F. Data quality

Problems:

- Big data contains many types of data, so it is not uniform and unbalanced [4].

- Big data collected from many sources, so the data quality is low and noisy [4].
- Data quality depends on big data technology [10].

Takes a long time to prepare data.

Solution: Blockchain technology with integrity, immutability, and smart contract limit noisy data and make it fully structured. Using blockchain technology to store big data can improve data quality and improve mining efficiency.

G. Data access performance

Problems:

- Multiple processes and departments involved in the data lifecycle [5].

Verification process needs to be done many times before having data access.

Solution: Through the blockchain network, blockchains encrypted and stored entirely on a distributed ledger. Thus, users can access data quickly and securely.

III. Blockchain services for big data

A. Collecting data

1) Secure data collection: Cloud computing widely applies to collecting big data from many sources. Therefore, it is difficult to control and ensure the safety of the big data collection process. Unofficial data sources and communication links are vulnerable vulnerabilities.

Solution: DRL framework [11] uses blockchain technology and reinforcement learning to solve this problem. The distributed blockchain helps to expand the scope of data collection from the terminals. Distributed ledgers and smart contracts ensure reliability and security as terminals share data without the need for a trusted third party. The process of collecting data from many sources is secure thanks to blockchain technology.

2) Transfer and share data securely: Big data needs to be analyzed and updated continuously over time from multiple data sources. The transmission and sharing of data from many sources easily make it vulnerable to cyber-attacks.

Solution: Chenhan et al. [12] propose a solution using blockchain technology to support data transmission and sharing. The authors use blockchain-based collaborative proof and useless transaction filtering algorithms to reduce processing time and storage costs. The process of transferring and sharing data is secure thanks to the distributed and immutable nature of blockchain.

B. Data storage

1) Secure file system: Some cloud services provide storage and accessibility anywhere for big data. Encryption methods implement to ensure the security of sensitive data. However, organizations and users still have security concerns when leaving the right to manage sensitive data to third parties.

Solution: To control the access to medical data (patient records, medical images, diagnostic reports, etc.) in the cloud storage, Sun et al. [13] integrate blockchain technology into the InterPlanetary File System (IPFS). Medical data encrypted based on the attribute and uses the private key associated with the attribute to access the data. The unique hash of the data used to authenticate the user. Blockchain distributed ledger used to record data storage and access. Blockchain immutability and transparency ensure the security of sensitive data.

2) Database management system: Data stored from various database management systems is vulnerable to attacks from both internal and external data sources. The method of using one-way hashing and digital marking is not applicable to distributed databases.

Solution: A solution applies blockchain technology at the proposed time stamp to ensure the security of the distributed database management system and the data transaction process (data storage, access, and sharing) between different database management systems [14]. All data transactions connected by a cryptographic hash function and stored on a common distributed virtual ledger. Smart contracts implemented to control the storage and sharing of data. Blockchain provides a secure transaction and storage environment for a distributed database management system.

C. Data analysis

Training data and building secure learning models: Big data formed from many data sources (social networks, websites, IoT devices, etc.) is the driving force behind the development of machine learning and artificial intelligence. However, if the collected data is fake, the quality of the machine learning model will decrease and lead to wrong decisions.

Solution: Solution for training data and building secure learning models from various data sources based on blockchain proposed by Shen et al. [15]. This is a method of training data and building an SVM model from multiple vertically partitioned data sources. Blockchain network and storage system applied homomorphic encryption combined to build a secure data training environment and build a secure learning model from a variety of data sources. Blockchain technology ensures a public, transparent and secure training data sharing environment.

D. Privacy Protection

Protect privacy: Big data continuously collects, updates, and examines personal data from various sources. As a result, big data can help organizations improve system quality and build better decision support systems. Organizations that manage big data benefit the most. However, issues of leakage of sensitive information and misuse of personal data are likely to arise.

Solution: Blockchain provides solutions to protect user privacy [16]. Blockchain technology, data encryption, and distributed ledgers ensure users didn't be identified, tracked, and monitored for sensitive personal information. Thanks to blockchain technology, users can easily access and track the transaction history (store, access, and share data) of their personal information.

IV. Blockchain big data applications

A. Smart city

- In the smart city environment, big data form from many facilities and sensor networks. These data streams transmit and store on cloud computing platforms. Third-Party Auditors (TPAs) are implemented to ensure data integrity and reliability. However, TPAs are a centralized entity, so they are vulnerable to attacks.
  Solution: [17] proposes a decentralized big data audit method for smart city. Based on blockchain technology, this method designs data auditing blockchain (DAB) to collect audit evidence and uses consensus algorithm. This approach eliminates centralized TPAs, which improves data reliability and integrity.
- The amount of data from user groups and IoT devices in smart cities will grow strongly in the coming time. This data needs to be stored, processed, and responded to quickly and with awareness. Specifically, the sharing economy in a smart city needs an environment to share data and conduct transactions safely and sustainably.
  Solution: [18] proposes a sharing economy system that can make transactions securely and does not need a central third agency to verify. This system is based on blockchain technology and mobile edge computing (MEC). In this system, data from IoT devices will be processed by MEC nodes. Key transactions are anonymous and stored securely on blockchain and off-chain systems. This system can be tested on a large scale in the near future.

B. Smart healthcare

- Remote patient monitoring to support healthcare and disease treatment is becoming more and more popular. IoT-based smart health monitoring devices are more commonly used by healthcare professionals. However, patient medical data is sensitive data. Therefore, there is a need for solutions to ensure patient privacy and data security during data processing.
  Solution: [19] proposes a new framework with a modified blockchain model suitable for the resource-constrained characteristics of IoT devices. This framework makes medical data control safer and more secure. The resource limitation of IoT devices and the cooperation of partners in the medical field are major challenges that need to be further considered and solved.
- In the medical field, sharing medical record data improves the quality of healthcare. However, this process faces many major challenges. Currently, there are many problems to ensure patient privacy and security in data sharing among medical stakeholders.
  Solution: [20] implements smart contracts in the healthcare sector. Based on the characteristics of blockchain technology, the system ensures that medical record sharing transactions are anonymous, safe, secure, and transparent. This solution helps to take advantage of available resources and has high system scalability.

C. Smart transportation

- Intelligent Transportation Systems (ITS) should ensure to provide accurate and complete traffic information. Malicious attacks and erroneous information will reduce system quality.
  Solution: [21] proposes to apply blockchain technology to the ITS system. Users can securely share data on the blockchain network and authenticate data sources easily.
- It is still challenging to ensure correct claim payment of transportation insurance services in big cities.
  Solution: [22] proposes a blockchain system that stores information about vehicle usage and driving behavior. The sensors on the vehicle collected this IoT information. Insurance services evaluate and pay claims based on this system's reliable data. Therefore, it helps make sure to pay compensation correctly. This solution can be widely deployed in large cities and applied in similar fields.

D. Smart grid

- The smart grid needs to ensure flexible data regulation during operation.
  Solution: [23] proposes a data aggregation and regulation mechanism based on blockchain technology. In addition, this method also uses encryption algorithms to collect multidimensional data. Thus, the smart grid can reduce operating costs, ensure security and analyze real-time data.
- The smart grid needs to ensure safety in the process of providing load management, frequency regulation, and grid stabilization services. Monitoring the entire operation of the smart grid to ensure safety is a big challenge.
  Solution: [24] introduces GUARDIAN management scheme based on blockchain technology. Energy transactions execute and store securely on the blockchain network automatically. Thus, data analysis is secure and helps to make safe energy business decisions.

V. Blockchain big data projects

A. Storj: Distributed storage system [25]

- Storj is a distributed network of computers that leases unused hard drive space from users to form a distributed storage system. This platform developed based on blockchain technology, using distributed ledger and encryption mechanism.
- The platform has 3 main components: storage node (user rents storage space), Uplink (launches at client machine to upload data to network), satellite (controls storage traffic and links archive buttons).
- Users participating in this system use Storj coin (ERC20 design standard, working on the Ethereum blockchain) to pay for transactions and pay for storage.

B. Omnilytics: Safe data trading platform [26]

- Omnilytics creates a secure data trading marketplace based on blockchain, big data analytics, machine learning, artificial intelligence, etc. This platform integrates data from various industries such as sales, marketing, and merchandise. Blockchain technology ensures a safe and secure way to contribute personal data, creating a common machine learning model.
- The way Omnilytics works is similar to a smart contract. The implementation of the smart contract will ensure that the payment will be made when participating in the construction of the common model.

C. Rubix: Distributed computing platform based on blockchain technology [27]

- Rubix has 2 types of tokens: Utility and Asset. The Rubix Chain consists of many parallel chains of accounts. Thus, cryptocurrency transactions can reach individual consensus and realize parallelization. This feature enhances the scalability of the system. So, Rubix's Proofchain architecture has some outstanding features compared to Ethereum. We can deploy safe, secure and highly scalable decentralized applications.
- The platform has the ability of integrating multiple e-traders into a common platform. From there, we can validate the reliability and predict the transaction more accurately.

D. Provenance: Open-source ecosystem for developing DeFi applications [28]

- Provenance is a public blockchain network that provides a ledger, a registry, and can perform many types of transactions through smart contracts. Transactions execute automatically through smart contracts on financial markets or supply chains.
- This platform supports building a supply chain that is traceable to information. Blockchain technology ensures transparent, safe, and fast information retrieval.

E. FileCoin: Distributed file system [29]

- FileCoin is an open-source project building the InterPlanetary File System (IPFS). The IPFS system is capable of linking different computing devices into the same peer-to-peer file system.
- FileCoin's decentralized storage method works similarly to Storj. However, Storj controls the transaction prices in their markets. FileCoin is community-driven, creating a price-competitive marketplace between tenants and lessors.

F. Datum: Distributed ecosystem for storing and exchanging data [30]

- Datum is a decentralized NoSQL database platform. It is decentralized and distributed using blockchain distributed ledger. Datum developed from Ethereum, Bigchain DB and IPFS.
- In Datum, any user can back up data from social networks, IoT devices, etc. safely, securely, and privately. This decentralized ecosystem allows users to share or sell data on their terms.

VI. Development challenges and future directions

Big data and blockchain are two technologies that have many characteristics that go well together. Therefore, these two technologies can integrate and support each other to improve performance. There are many surveys conducted to assess the extent and scope of integration of these two technologies. Currently, proposed solutions generally focus on four problems: security and privacy in big data, big data management and exploitation, blockchain standardization, and network virtualization. Detailed content about integrated solutions and oriented development of these problems presented in Table I.

However, there are many challenges when combining these two technologies. Some challenges when applying blockchain technology to real projects are: computational complexity increases as the blockchain length increases, the complexity of data stored in the blockchain is increasing, ensuring system performance, etc. Existing blockchain big data projects continue to upgrade and find solutions to these problems.

VII. Conclusions

Big data applied more and more widely. The growth rate of this market is very fast. This field attracts the special attention of researchers. Research projects on opportunities and solutions to develop big data are increasing day by day. In this article, we focus on analyzing the studies of integrating blockchain technology into big data to solve some important problems in big data. We presented the basis for integrating these two technologies, blockchain solutions for big data storage cloud services, application scope of blockchain for big data solutions, big data blockchain projects, and direction to develop in the future. We surveyed and presented a complete picture of a blockchain solution for big data. This helps researchers identify the problem and continue to develop it in the future.

Acknowledgment

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number C2022-26-05.

References

[1] Khan, N. A. (2019). The 51 v's of big data: survey, technologies, characteristics, opportunities, issues and challenges. Proceedings of the international conference on omni-layer intelligent systems.

[2] Zhou, Q. H. (2020). Solutions to scalability of blockchain: A survey. Ieee Access, 8, 16440-16455.

[3] Tang, M. M. (2017). Big data for cybersecurity: Vulnerability disclosure trends and dependencies. IEEE Transactions on Big Data 5, 3, 317-329.

[4] Deepa, N. Q.-V. (2022). A survey on blockchain for big data: approaches, opportunities, and future directions. Future Generation Computer Systems.

TABLE I
STATISTICS OF DEVELOPMENT CHALLENGES AND FUTURE DIRECTIONS

Challenges	Integrated solution	Oriented development
Security and privacy in big data	[31] explores the use of smart contracts to perform data sharing transactions in big data. Smart contracts execute trusted transactions between multiple objects without the need for a third party.	Learn the security issues of smart contracts and the correct programming method for smart contract verification. Smart contract in blockchain technology is a promising solution to share data securely and ensure privacy in big data.
Big data management and exploitation	[32] explores the convergence of big data, blockchain and cryptocurrency. Integrate blockchain technology for decentralized data management and mining.	Identify problems and develop solutions to integrate blockchain technology into big data management and exploitation. Applying blockchain technology to build decentralized big data management, query, and mining solutions.
Blockchain standardization	[33] explores the boundaries of immutability in blockchain technology. The framework proposed as a standard for implementing blockchain systems.	The proposed results' framework presents different levels of standardization and their benefits. As a result, the immutability of blockchain technology guarantees. Implement standardization early to harness the full potential of blockchain technology into big data.
Network virtualization	[34] explores how to combine SDN, big data, blockchain and 5G MEC in smart city. In addition, they discuss security issues of data exchange in smart city.	Deploy SDN and 5G MEC to speed up transmission and query big data in smart city. Implement blockchain technology into SDN architecture to ensure immutable and secure data transactions.

[5] Altulyan M, Y. L. (2020). A unified framework for data integrity protection in people-centric smart cities. Multimedia Tools and Applications, 79, 7, 4989-5002.

[6] Liu B, Y. X. (2017). Blockchain based data integrity service framework for iot data. 2017 IEEE international conference on Web services (ICWS).

[7] Jha, B. K. (2020). Fraud detection and prevention by using big data analytics. Fourth international conference on computing methodologies and communication (ICCMC).

[8] Rong, M. D. (2019). Feature selection and its use in big data: challenges, methods, and trends. IEEE Access, 7, 19709-19725.

[9] Dong, X. R. (2015). Secure sensitive data sharing on a big data platform. Tsinghua science and technology 20.

[10] Becker, D. T. (2015). Big data, big data quality problem. IEEE International Conference on Big Data (Big Data).

[11] Liu, C. H. (2018). Blockchain-enabled data collection and sharing for industrial IoT with deep reinforcement learning. IEEE Transactions on Industrial Informatics.

[12] Xu, C. W. (2018). Making big data open in edges: A resource-efficient blockchain-based approach. IEEE Transactions on Parallel and Distributed Systems.

[13] Sun, J. Y. (2020). Blockchain-based secure storage and access scheme for electronic medical records in IPFS. IEEE Access, 8, 59389-59401.

[14] Li, H. a. (2019). EduRSS: A blockchain-based educational records secure storage and sharing scheme. IEEE Access, 7, 179273-179289.

[15] Shen, M. Z. (2019). Secure SVM training over vertically-partitioned datasets using consortium blockchain for vehicular social networks. IEEE Transactions on Vehicular Technology, 6, 5773-5783.

[16] Bernabe, J. B.-R. (2019). Privacy-preserving solutions for blockchain: Review and challenges. IEEE Access, 7, 164908-164940.

[17] Yu, H. Z. (2018). Decentralized big data auditing for smart city environments leveraging blockchain technology. IEEE Access, 7, 6288-6296.

[18] Rahman, M. A. (2019). Blockchain and IoT-based cognitive edge framework for sharing economy services in a smart city. IEEE Access, 7, 18611-18621.

[19] Dwivedi, A. D. (2019). A decentralized privacy-preserving healthcare blockchain for IoT. Sensors, 19, 326.

[20] Vyas, J. D. (2020). Integrating blockchain technology into healthcare. Proceedings of the 2020 ACM Southeast Conference.

[21] Hîrţan, L. A.-V. (2020). Blockchain-based reputation for intelligent transportation systems. Sensors, 20, 791.

[22] Li, Z. X. (2018). Blockchain and IoT data analytics for fine-grained transportation insurance. IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS).

[23] Fan, M. a. (2019). Consortium blockchain based data aggregation and regulation mechanism for smart grid. IEEE Access, 7, 35929-35940.

[24] Jindal, A. A. (2019). GUARDIAN: Blockchain-based secure demand response management in smart grid system. IEEE Transactions on Services Computing, 13, 613-624.

[25] Zhang, X. J. (2019). Frameup: An incriminatory attack on Storj: A peer to peer blockchain enabled distributed storage system. Digital Investigation, 29, 28-42.

[26] Liang, J. S. (2021). OmniLytics: A Blockchain-based Secure Data Market for Decentralized Machine Learning. International Workshop on Federated Learning for User Privacy and Data Confidentiality in Conjunction with ICML.

[27] Bonsón, E. a. (2019). Blockchain and its implications for accounting and auditing. Meditari Accountancy Research, 27(5), 725-740.

[28] Kim, H. M. (2018). Toward an ontology-driven blockchain design for supply-chain provenance. Intelligent Systems in Accounting, Finance and Management, 25(1), 18-27.

[29] Psaras, Y. a. (2020). The interplanetary file system and the filecoin network. 50th Annual IEEE-IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S).

[30] Katiyar, D. a. (2021). Blockchain technology in management of clinical trials: a review of its applications, regulatory concerns and challenges. Materials Today: Proceedings, 47, 198-206.

[31] Liu, J. a. (2019). A survey on security verification of blockchain smart contracts. IEEE Access, 7, 77894-77904.

[32] Hassani, H. X. (2018). Big-crypto: Big data, blockchain and cryptocurrency. Big Data and Cognitive Computing, 4, 34.

[33] Hofmann, F. S.-S. (2017). The immutability concept of blockchains and benefits of early standardization. 2017 ITU Kaleidoscope: Challenges for a Data-Driven Society (ITU K).

[34] Aujla, G. S. (2020). Blocksdn: Blockchain-as-a-service for software defined networking in smart city applications. IEEE Network, 34, 83-91.