Blockchain : A Decentralized Approach to Big Data

Senem Pehlivan Kaplan[1] and Assist. Prof. Dr. Serap Şahin[2]

[1] Sirnak University, Sirnak/Turkey, senemkaplan@sirnak.edu.tr
[2]Izmir Institute of Technology, Izmir/Turkey, serapsahin@iyte.edu.tr

***Abstract* - The blockchain technology is a hot topic and a new technology in recent years. It is not only an underlying technology for many applications like Bitcoin application, but also it is a kind of thinking including cognitive and mental processing and understanding for artificial intelligence and human enhancement. All data or services are digitized. So, this leads to deal with big data. It is a challenge to deal with big data from the perspective of performance, scalability, availability and privacy in centralized systems. Blockchain is applicable to big data and brings different perspective how to process, store, read and write data. Also, the aim of this paper is to show that better solutions are possible in a decentralized way. Even the technology is in its early stages, the blockchain technology will be in future due to its superior features. Therefore, it is better to adopt this technology as soon as possible to place in future. This paper gives a brief about how the blockchain could approach to big data and analyzes existing information regarding the challenges of big data from the side of the blockchain.**

***Keywords* – Blockchain, big data, decentralized systems, scalability, performance, availability, privacy.**

I. INTRODUCTION

Today, data is growing larger rapidly, since everything is digitized. Data could be structured or unstructured. Dealing and processing big data is a big challenge without sacrificing security, performance, scalability and availability. First applications to come to mind are social networks, IoT, mobile apps, healtcare etc. using videos, images, documents, electronic records, medical measurements. Also, these applications based on centralized client-server model has some limitations and concerns.

In recent years, a new technology called blockchain has drawn lots of attention and got much interest with its most popular and known application named Bitcoin. The first appearance of the blockchain is when Satoshi Nakamato comes up with a digital currency called Bitcoin using blockchain as an underlying technology in his white paper [1] in 2008.

Blockchain (BC) is a shared, distributed and decentralized ledger technology. BC uses P2P communications. Moreover, there is no trusted third party in the blockchain. It is an underlying technology for broad range of applications. Moreover, [12] emphasizes that BC is not only underlying technology, but also it is a kind of thinking for machines and humans. In BC thinking, processing occurs in a decentralized way.

The blockchain is categorized in three types such as public – permissionless, private – permissioned and consortium BCs in [5]. Bitcoin and Ethereum are the examples of the public ones. HyperLedger Fabric is an example of the private one. In the public BC, any node can join and leave the network. There is no restriction about it. On the other hand, in the private BC, to access the network, permission is required. If any node requests to join the network, it must be approved from other nodes in the BC private network. The study of [10] claims that this allows for accountability and compliance with laws and regulation and also supports openness and collaboration with participants. In consortium BC, preselected nodes determines the consensus process [5].

In BC, each transaction is stored in a block. After the transaction and block is verified, it is chained with other blocks. So, each node in the blockchain network has a copy of whole transactions, which provides transparency. Each record is stored permanently. In the private blockchain, only verified nodes can join the network. Because any trusted authority doesn't exist in BC, trust concept is based on cryptographic proof. So, this reduces costs and provides easiness. Even it is in early stages, it has already been used in many areas. Banking and finance, government processes, social media, healthcare, education operations, notary, real estate, IoT (Internet of Things) etc. Its approach is also applicable to big data and brings better solutions to handle big data.

Processing big data and keep sensitive information secure is a concern even in centralized systems. If the centralized system crashes, it could lead to data loss. Even GFS (Google File System) is based on the idea that component failures are the norm rather than exception [16]. This prevents data loss and guarantees availability. In the blockchain, all records is stored in all nodes in the network. Because the number of nodes are thousands or millions, and even some nodes crash in the network, data will be available all the time. However, storing any data in all nodes could lead to some storage costs or capability problems, which are still researched and there exists different type of solutions about these issues. Because there is no third party in the blockchain, transactions are done in roughly ten minutes, while they take a few days in centralized systems due to being intermediary in between.

It is possible to store all transactions including big size data like videos, pictures etc. Scalability is an issue in the blockchain while data gets larger. There are different studies and approaches to big data from the blockchain perspective.

In the blockchain network, anything can be exchanged. It could be currency, goods or service such as user post in social

networks, digital currency in banking, service in government agencies etc.

In existing systems, there are double-spending problem in banking and finance, higher costs not to stay anonymous or privacy issues from users' side. To overcome these, BC could be an alternative solution.

We know the blockchain technology is in its early stages. Scalability and security issues are still on table. To specify and draw its usage and range of application is not known yet, for it is the technology could applied to whole network, or mediator, public or private or even hybrid, we can extend or narrow it according the our application. Therefore, we need to consider this technology from the wider perspective due to its flexible feature.

In the next section, the important terms on the blockchain are described. In the third section, the limitations of centralized systems are reviewed. In the fourth section, pros and cons of the blockchain and its solutions for big data and, also, the existing applications of blockchain are analyzed and examined.

II. IMPORTANT TERMS IN BC

This section includes the definitions of blockchain related terms and concepts. These are important to explain the blockchain domain.

Node — It is a computer in the blockchain network.

Miner — A special node is usually rewarded to mine on the BC network.

Block — A block is a structure where each record is stored. The blocks are timestamped and then are chained in the blockchain. Each block consists of the block header, merkle tree and previous block's hash.

Distributed Systems — It consists of collection of independent computers. Whole system behaves like a single coherent system. The idea of BC is based on distributed systems.

Distributed ledgers — All records are shared and replicated among all nodes in the blockchain network.

Fork — A block could reference two or more blocks. Any user of the Bitcoin system may maintain a local copy of the blockchain and resolves conflicts by believing in the longest chain [14].

Consensus Process — In BC, each transaction or record must be reviewed, since after verifying transactions, they are immutable and irreversible. Also, there is no central authority in BC. Therefore, consensus is necessary and there are some consensus algorithms we will review here.

Consensus algorithms :

1. *Proof of Work (PoW)* : A block is hashed with a nonce in order to get a hash with zero bits at the beginning of hash. This consensus algorithm could use SHA-256 as used in [1]. Proof of Work effort is dependent on energy consumption and CPU power.

2. *Proof of Stake (PoS)* : This is an alternative consensus algorithm of Proof of Work. In PoS, mining is dependent on the amount of digital assets each node has.

3. *Proof of Activity* : This is a consensus algorithm of hybrid of Proof of Work and Proof of Stake.

4. *Proof of Intelligence* : It is a consensus algorithm dependent on reputational asset on the blockchain thinking.

In addition to these most known consensus algorithms, there are other consensus algorithms such as Proof of Importance, Proof of Authority etc., which we will not mention in this paper.

Mining — It is a process to make transactions valid at the expense of computational power. In Bitcoin BC, on an average, every 10 minutes, a new block is appended to the blockchain through mining [15].

Incentive — Mining is rewarded. For example, in Bitcoin, it is rewarded with coins. The reward is 12.5 BTC at the time of writing. Also, in [9], MedRec, healthcare management system based on BC, the reward is to access aggregate, anonymized medical data. Another example is that [6] representing a traffic announcement system has a reward mechanism informing other drivers with announcement of the accident to earn some coins.

Smart Contract — It is a small computer code which is executed on the blockchain. Ethereum blockchain nodes are capable enough to execute any kind of smart contracts.

III. LIMITATIONS OF CENTRALIZED SYSTEMS

The existing systems mostly adopt centralized client-server model. However, some businesses change their system to distributed ones. Distributed systems solve some issues, but it is not enough without combining both distributed and decentralized systems. The reason for transferring data to distributed systems is why centralized systems even looks like better to control of data, focus on data and be more consistent, and also it is not flexible and brings some issues to process a large variety and amount of data sets.

For most of applications, it is critical to access data all the time such as banking applications, government processes etc.. Depending on the centralized system has risks at crashing servers affecting whole system unavailable. This could be a disaster for banks and users.

Privacy is another concern from the side of centralized systems. Social networks gather all personal data of users and user preferences. Users has no control of how their data is processed and at whose hands. The privacy of users is violated due to imposing irreversible user preferences. The same problem occurs at in different area called Internet of Things (IoT). IoT objects has direct connections between users and IoT object providers. These providers gather all different kind of personal data including actions and habits of users. In addition to these, mobile applications has privacy issues for users. Once a user accepts a privacy policy of a mobile

application by installing it, it is unchangeable unless the mobile app is uninstalled.

In centralized systems, while large data sets grow exponentially, scalability and performance issues come up. Moreover, servers have permitted capability. Therefore, exceeding this capability leads to degrade performance to respond clients and Denial of Service (DoS). That is, these systems have scalability bottlenecks. For example, we can tackle healthcare systems. Healthcare systems deal with large size of data such as imaging, documents, electronic records, etc. The number of patients and their documents increases and their capability is limited. So, existing healthcare systems, have scalability concerns and, also, are not interoperable and convenience because different hospitals can not see their medical records, which is highly vital to track of patient's medical history and provide fast feedback to patients.

IV. BIG DATA ON BLOCKCHA IN

How BC approaches big data is a key concept to shape both today and future. BC has the distributed and decentralized architecture. It is immutable and irreversible. In BC network, any node has a whole copy of ledger. There will be thousands or millions nodes. Even some of computers crash in BC network, data loss is prevented and data is kept available by this way, since the decentralized nature of the BC is fault-tolerant. Trusted third party models such as banking, social networks are unavailable for security and maintenance purposes. BC enhances the availability of data while comparing it to traditional centralized systems by eliminating third parties. For example, file not found or server unavailable error messages will be mitigated at a significant percentage.

While we tackle social media networks, Internet of Things, mobile applications or data or digital content ownership applications suffer from especially privacy due to centralized systems. The existing applications don't guarantee user privacy. User has a lack of control of their data. The reference [4] proposes a user-centric permissioned blockchain based social media network called Ushare. The consensus algorithm of Ushare is Proof of Stake to validate transactions and blocks. As pointed out in [4], Ushare provides more secure social media network, for content is encrypted and Ushare works on a distributed hash table similar structure Bigtable [17]. Bigtable is established as distributed storage system managing structured data for large size data in order to achieve better scalability and higher performance used by Google as stated in [17]. A Personal Certificate Authority (PCA) is introduced to allow only user's circle members to see the user shared content [4].

It is obvious that the number of IoT objects are increasing every day. This brings privacy, scalability and performance concerns with it. Therefore, first of all, it is significant to provide privacy-aware solutions. In [7], a solution is brought for the privacy of IoT and mobile applications with the blockchain. In this design and implementation, the blockchain is an intermediary between users and IoT application or mobile application providers for management of privacy preferences. In the Ethereum blockchain, privacy policies and device's info are embedded in the smart contracts. Accepting and declining privacy policies are done via blockchain connected gateway. If the user accepts the privacy policy, whenever he connects IoT devices through the blockchain, the blockchain network preserves his privacy preferences. The digital signature scheme is based on ECDLP (Elliptic Curve Discrete Logarithm Problem). All user preferences are encrypted and stored in the blockchain network. The user can connect the gateway and review the privacy policies and information of the device. The computational cost of PDSS (Proposed Digital Signature Scheme) is nearly 283 ms considered practical and reasonable in real-life applications as pointed out in [7]. First implementation of Internet of Things based on the Ethereum blockchain network is Slock.it.

One of the area managing big data is healthcare. MedRec is a healthcare application based on BC. MedRec targets to access faster to medical data, to overcome interoperability; to enhance medical research with patient's data. MedRec's aim is to achieve security, privacy, interoperability and scalability issues by keeping immutable logs related to patients. In private blockchain networks especially, some modifications can be made for higher scalability and performance [9]. For this reason, MedRec is based on private, Ethereum BC which utilizes smart contracts categorized in registrar contract, patient-provider relationship contract and summary contract.

There are scalability and performance bottlenecks in the blockchain. While comparing transaction rate of visa and bitcoin, the most common application of BC, it is seen that Visa can process 4000 transactions per second, on the other hand, bitcoin processes 7 transactions per second. However, as a first step, re-parameterization of block size for 4 megabyte in Bitcoin and block intervals for roughly 10 minutes in Bitcoin make difference to get higher throughputs and lower latencies [19]. In addition to these, adopting different protocols will definitely increase scalability and performance of whole broad range of blockchain applications by maintaining the decentralization architecture of the blockchain. The abstraction layers in [19] are defined as network layer, consensus layer, storage layer, view layer and side layer, respectively from bottom to top. To analyze each layer leads us to find solutions to achieve scalability and performance issues. It is emphasized that broadcasting messages are made in network layer and also after propagating transactions to the BC network, a block is propagated. This means that transactions are stored in the block structure, so transactions transmitted twice. In consensus layer, different protocols such as Byzantine Fault Tolerant protocols or Paxos could be used but sharding protocols seem to be necessary to improve scalability of the BC. After authenticated data in consensus layer, in addition to data, smart contracts and views are stored and available in storage layer, which is vital to be developed for growing data [19].

EduCTX is also a consortium blockchain-based platform and works on delegated proof of stake (DPoS) as a consensus algorithm. Its aim is to transfer credits and grades among higher education institutes globally. This platform is

not restricted to any certain languages thanks to smart contracts. In this consortium blockchain, only higher education institutes can join the network after creating a blockchain wallet and verifying by the other nodes. Each student has a blockchain address and then, each ECTX tokens regarding the students' completed courses are transferred to the students' blockchain addresses.

Hyperledger Fabric is mostly known as a cryptocurrency, but actually it is an permissioned blockchain supported by IBM in order to be utilized as an underlying technology for a broad range of area such as finance, IoT, healthcare etc. The reference [10] emphasizes that in Hyperledger Fabric blockchain network, performance optimizations depends on code generated by compiler, the performance of cryptographic algorithms and enhancements specific. When an asymmetric algorithm is established on elliptic curves increases performance significantly.

V. CONCLUSION

To deal with big data, it is important to achieve certain issues such as privacy, availability, scalability and performance. At this point, understanding BC provides alternative and better solutions than existing traditional systems. Analyzing and summarizing applications built based on the BC will give us insight to evaluate and propose different approaches. Retaining the decentralization feature of BC, it is possible to choose the corresponded type of the blockchain and consensus algorithm for different areas.

REFERENCES

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system", White Paper, 2008.

[2] N. Bozic, G. Pujolle and S. Secci, "A tutorial on blockchain and applications to secure network control-planes," in Smart Cloud Networks & Systems (SCNS), Dubai, UAE, December 2016.

[3] E. Karafiloski, A. Mishev. "Blockchain solutions for big data challenges: A literature review." In: IEEE EUROCON 2017 -17th International Conference on Smart Technologies. July 2017, pp. 763–768. DOI: 10.1109/EUROCON.2017. 8011213 (cit. on p. 16).

[4] A. Chakravorty, C. Rong, "Ushare: user controlled social media based on blockchain", International Conference on Ubiquitous Information Management and Communication, 2017.

[5] M. Turkanovic, M. Hölbl, K. Kosic, M. Hericko, A. Kamisalic, EduCTX: A blockchain-based higher education credit platform. IEEE Access 2018, doi:10.1109/ACCESS.2018.2789929.

[6] L. Li. , J. Liu, L. Cheng, S. Qiu, W. Wang, X. Zhang., Z. Zhang, "CreditCoin: A Privacy-Preserving Blockchain-Based Incentive Announcement Network for Communications of Smart Vehicles", IEEE Transactions on Intelligent Transportation Systems. PP. 1-17. 10.1109/TITS.2017.2777990, 2018.

[7] S. Cha, J. Chen, C. Su, K. Yeh, "A Blockchain Connected Gateway for BLE-based Devices in the Internet of Things",IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2799942.

[8] G. Zyskind, O. Nathan and A. '. Pentland, "Decentralizing Privacy: Using Blockchain to Protect Personal Data,"2015 IEEE Security and Privacy Workshops, San Jose, CA, 2015, pp. 180-184. doi: 10.1109/SPW.2015.27.

[9] A. Ekblaw, A. Azaria, J. D. Halamka, MD, A. Lippman,n"A Case Study for Blockchain in Healthcare: "MedRec" prototype for electronic health records and medical research data", White Paper, 2016.

[10] N. Mencias, D. Dillenberger, P. Novotny, F. Toth, T. E. Morris, Jr., V. Paprotski, J. Dayka, T. Visegrady, B. O'Farrell, J. Lang, E. Carbarnes. (2018). An optimized blockchain solution for the IBM z14. IBM Journal of Research and Development. PP. 1-1. 10.1147/JRD.2018.2795889.

[11] A. Dorri, M. Steger, S. S. Kanhere and R. Jurdak , "BlockChain: A distributed solution to automotive security and privacy, [online] Available: https://arxiv.org/abs/1704.00073.

[12] Swan M, "Blockchain thinking: the brain as a decentralized autonomous corporation," IEEE Technology and Society Magazine, vol. 34, no. 4, pp. 41-52, December, 2015.

[13] S. Fujimura, H. Watanabe, A. Nakadaira, T. Yamada, A. Akutsu, J.J. Kishigami, "Bright: A concept for a decentralized rights management system based on blockchain", 2015 IEEE 5th International Conference on Consumer Electronics — Berlin (ICCE-Berlin), pp. 345-346, Sept 2015.

[14] M. Moser, R. Bohme, D. Breuker, "An inquiry into money laundering tools in the Bitcoin ecosystem", IEEE eCrime Researchers Summit (eCRS), 2013.

[15] G. Foroglou, A. L. Tsilidou, "Further applications of the blockchain", 2015.

[16] S. Ghemawat , H. Gobioff , S. Leung, "The Google file system", Proceedings of the nineteenth ACM symposium on Operating systems principles, October 19-22, 2003.

[17] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, R.E. Gruber, "Bigtable: a distributed storage system for structured data", 2006.

[18] C. Jentzsch,"Decentralized Autonomous Organization to Automate Governance", White Paper, 2015.

[19] K. Croman, C. Decker, I. Eval, A. E. Gencer, A. Juels, A. Kosba, A. Miller, P. Saaxena, E. Shi, E. G. Sirer, D. Song, R. Wattenhofer, "On Scaling Decentralized Blockchains", (A Position Paper). In 3rd Workshop on Bitcoin and Blockchain Research, 2016.