SOLVING PROBLEMS OF BIG DATA INFRASTRUCTURE BY USING BLOCKCHAIN

MOHAMED ESMAIL[1], MOHAMED A. EL-DOSUKY [1, 2], TAHER T. HAMZA[1]

[1] Faculty of Computers and Information Sciences, Mansoura University, 35516, Egypt

[2] Dept. of Computer Science, Arab East Colleges, P.O.Box 53354 Riyadh, Saudi Arabia

E-mail: [1] mesmail922@gmail.com

ABSTRACT

Big data faces challenges like substructure security, data confidentiality and data administration. With the inception of blockchain, massive data security circulation has become possible. The paper surveys previous work, focusing on security challenges in big data models, a summary of blockchain services in a big data environment, and the challenges of research in big data working together with blockchain. Proposed methodology begins with the mathematical foundation of big data-blockchain mapping. Then the proposed big data blockchain infrastructure is proposed. The infrastructure allows many operations such as separating and storing data, querying the separated stored data, block validation, and gossip protocol. This work established an archetype system using python programming language to confirm that the suggested concept of isolating and information storage is applicative to big data controlling systems on blockchain technologies. Finally, forged block attack is scrutinized. The methodology integrates big data with blockchain through a proposed infrastructure, validated via a Python prototype, enabling data management and addressing forged block attacks.

Keywords: *Big data, Blockchain, Security*

1. INTRODUCTION

United States, was conquered with the "DDoS attack" [1], which led to countless terror in the safety of information in October 2016. Then the world confirmed the existence of the shortcomings of central storing of data assets [2]. Same year, EU admitted General Data Protection Regulations (GDPR) [3]. As well as the China Academy of Information and Communications Technology (CAICT) released the Data Safety Transmission Agreement. All protocols cheer standardization of big data processing, preserves data transmission to a limit grade. But despite that, governmental protocols didn't control people's attitude, issue basically not resolved. In contrast with the newcomer of the blockchain [4], resolving security flow has come to be probable. Blockchain produces smart contracts notion [5], no need to deal among currencies. Smart contracts execute code, therefore it does not demand reciprocal confidence, over and above, a hundred percent programmed and never tampered with.

The motivation behind this work is to address the challenges of big data management by leveraging blockchain technology to enhance security, integrity, and scalability. With the increasing reliance on decentralized systems, blockchain provides an immutable and transparent solution for storing and querying massive datasets, preventing data manipulation, and ensuring trust across distributed networks. This work explores the integration of blockchain with big data systems to offer a secure, efficient, and scalable approach, while also tackling issues like forged block attacks and inefficient querying. Ultimately, it aims to demonstrate the feasibility of blockchain-based big data management through a prototype, providing a foundation for real-world applications in industries requiring secure and decentralized data storage.

This work focuses on integrating blockchain technology with big data management systems, specifically addressing the mapping of big data onto blockchain, designing a blockchain infrastructure for secure data storage and querying, and developing a Python-based prototype to validate the proposed system. The research also examines the impact of forged block attacks on data integrity and explores potential countermeasures. The scope is limited to enhancing the security, scalability, and efficiency of big data management using blockchain, without extending into other blockchain applications such as cryptocurrencies or smart contracts.

The remainder of the paper is ordered in the following way. Section 2 is for previous work, focusing on security challenges in big data, a general view on blockchain services among big data environment also challenges faced blockchain working together with big data. Section 3 is the proposed methodology, beginning with the mathematical foundation of big data-blockchain mapping. Then the proposed big data blockchain infrastructure is given. The infrastructure allows many operations such as separating and storing data, querying the separated stored data, block validation, and gossip protocol. Section 4 is for implementation and results. Section 5 is for conclusion and future directions.

2. PREVIOUS WORK

In the big data age: IoT, Social Media, and sensors cause data size to burst. In sum DAU (daily active users) entree applications owned by Facebook reach 2.91 billion containing Instagram, WhatsApp, and Messenger [6]. Lots of challenges facing big data. The major challenges are [7]: substructure security, data confidentiality, data administration, and data veracity.

2.1 Big Data Security Challenges

There are a lot of challenges facing big data. The major challenges are [18]: security of substructure, data confidentiality, integrity and administration of data, and reactive security.

2.2 Blockchain

Blockchain 2.0 era has arrived [8], Blockchain solved the massive data security circulation. Blockchain produces smart contracts notion [5], no need to deal among currencies. Smart contracts execute code, therefore it does not demand reciprocal confidence, over and above, a hundred percent programmed and never tampered with. By creating a blockchain storage model, the smart contract mechanically establishes an authoritative big data Circulation method without confidential third parties.

2.3 Blockchain Big data

Blockchain consists of network which is decentralized whereby the collaborating individuals have full right to watch communications of blockchain network P2P technique [11, 12]. Through 1600 crypto-currencies Bitcoin is the greatest public platform [13]. Finally, all individuals participate in transaction equal replica [14].

2.4 Acquisition of Big Data by Blockchain

Big data can be categorized as shapeless or organized. Organized data involves info already administered by the enterprise databases. Shapeless data is disorderly information from diverse sources which do not have a determined form. These data must be converted to a structured form to gain numerous foretelling. Blockchain uses consensus algorithms which ensures data integrity that's why attacks decreased. Smart contracts permit the trading of data by sharing idle data of customers, the recycled block joined to the blockchain network like a fresh node. Therefore, valuable forecasts are engendered in AI models.

2.5 Blockchain & Secure Database Management

Data stored in different database management systems; therefore, it is under offensives from inner and outer sources. Cryptography hash functions use database fraud exposure methods to discover malicious hacks in the databases. Blockchain techniques use time-stamping to avert data tampering by recording all transaction in collection of blocks. If attacker forged a block, forged block hash value reorganized. Figure 3 blockchain services propose a pattern store and share records securely. This pattern assimilates servers, cryptographic algorithms, and block chains to progress trustworthy surroundings [18].

As mentioned in Table1 Blockchains three categories (Public, Private and Consortium). This is a model for Big Data blockchain layer which shows visualization, blockchain and monitoring layer.

Proof of Work protocol steps are as follows:

- Bunch transactions into blocks.
- Miners verify the legality of each transaction.
- Miners have to find a solution to a mathematical puzzle known as a proof-of-work issue.
- A reward is given to the first miner who reach solution for the puzzle.
- Then the verified transactions are saved in the public blockchain.

3. PROPOSED METHODOLOGY

First, assume the arithmetic basis of big data-Blockchain. It is inspired from IOT-Blockchain mapping.

Theorem 1 (Big data-Blockchain Mapping)

$$\lambda_{BC} \leftarrow \frac{\delta_{BC}(\Phi_{BD}-\lambda_{BD})}{\Phi_{BD}} + \theta_{Bc} \qquad (1)$$

where λ_{BD} is the feature vector of big data, such as (volume, variety, velocity). $\lambda_{BD} \in [0, \Phi_{BD}]$.

and λ_{BC} is the feature vector of blockchain as in **Table 3** . $\lambda_{BC} \in [\theta_{Bc}, \theta_{Bc} + \delta_{Bc}]$.

Proof

λ_{BD} is normalized as follows:

$$0 \le \lambda_{BD} \le \Phi_{BD} \quad (2)$$

$$0 \le -\lambda_{BD} \le -\Phi_{BD} \quad (3)$$

$$\Phi_{BD} \ge \Phi_{BD} - \lambda_{BD} \ge 0 \quad (4)$$

$$1 \ge \frac{\Phi_{BD}-\lambda_{BD}}{\Phi_{BD}} \ge 0 \quad (5)$$

$$0 \le \frac{\Phi_{BD}-\lambda_{BD}}{\Phi_{BD}} \le 1 \quad (6)$$

By multiply blockchain feature length

$$0 \le \frac{\delta_{BC}(\Phi_{BD}-\lambda_{BD})}{\Phi_{BD}} \le \delta_{Bc} \quad (7)$$

By adding minimum value of blockchain feature

$$\theta_{Bc} \le \frac{\delta_{BC}(\Phi_{BD}-\lambda_{BD})}{\Phi_{BD}} + \theta_{Bc} \le \theta_{Bc} + \delta_{Bc} \quad (8)$$

By this

$$\frac{\delta_{BC}(\Phi_{BD}-\lambda_{BD})}{\Phi_{BD}} + \theta_{Bc} \quad (9)$$

λ_{BC} is a blockchain feature in interval $[\theta_{Bc}, \theta_{Bc} + \delta_{Bc}]$. Denoting it λ_{BC}

$$\lambda_{BC} \leftarrow \frac{\delta_{BC}(\Phi_{BD}-\lambda_{BD})}{\Phi_{BD}} + \theta_{Bc} \quad (10)$$

As shown in Fig 6, MVC model which shows Model, View, and Controller Layer is presented. In Model layer HDFS data is presented then distribute them into blocks through Gossip Protocol. Controller layer managed the CRUD functions. In View layer shows the user interface results.

Despite of sharing data between organizations and each other is a big problem, Consortium Blockchain allow various trusted enterprises to share data in the network. Consortium Blockchain uses tamper-proof safekeeping characteristic which decrease enterprises distribution data worry, our suggested on/out-of-chain prototypical stock a huge data in local databases, so it is conventional for traditional enterprises.

Based on Theorem 1 The Block validation algorithm in this archetype steps:

- Make sure former block of existing block found and valid.
- Ensure block timestamp is bigger than former block timestamp and lower than 2 hours in the future.
- Make sure proof of work of block is valid.
- Suppose BS[0] be former block's finish status.
- assume we have list n transactions; Tr is transactions of the block for all j in 0...n-1, set BS[j+1] = APPLY (BS [j], Tr [j]), if any scenario returns mistake: exit, and return false.
- Return true, and record BS[n] to be end status of this block.

Network launches with one parent node then a collection of energetic nodes. It's necessary for every single node to have one neighbor at least. thus, when a fresh node releases, Gossip script chooses a randomly energetic neighbor node. Fresh node demands current node to become its neighbor. Received nodes add fresh node to neighbor's list then send notification to the fresh node (see Fig.10). With every fresh node acts, the script reiterated to compose a randomly network of joined nodes

4. IMPLEMENTATION AND RESULTS

This work established an archetype system using python programming language to confirm that the suggested concept of isolating and information storage is applicative to big data controlling systems on blockchain technologies. Hadoop completely stretchy, which make capacity planning easier for clusters. Hadoop cluster used for the spread computing, where it can store and analyze massive amounts of structured and unstructured data

4.1 Formula to compute HDFS nodes Storage

H = C*R*S/(1-i) * 120% (11)

Where:

C = Compression(Solidity) ratio.

R = Replication factor.

S = Data Size needs to stir to Hadoop.

i = midway data factor.

When there is no solidity: C = 1, R = 3, and midway factor of 0.25 = 1/4

H = 1*3*S/(1-1/4) = 3*S/(3/4) = 4*S

Hadoop storage is evaluated to be 4 intervals the size of the original data size after above presumptions

Let's assume the following:

Daily Ingestion rate 1 TB

Replication Factor 3

Size of Hard Disk 48 (12 * 4 TB)

Buffer memory 25% or 0.25

Memory to be stored in HD 1 * 3 = 3TB

Memory can be used for storing and processing 48-(48*0.25) = 36 TB

Number of Nodes required (3*365)/36 =~31 Nodes

If we assume that a cluster can store and analyze 36TB and we have 100 cluster, then the maximum volume of our data will be 360 TB.

4.2 Forged Block Attack Control

Blockchain is warranty of big data safety issues. Let's assume an attacker forged a node in the blockchain network, then faking blocks to connect. the honest chain competes the attackers chain by walking randomly in a binary hierarchy [39]. We can calculate the of attack success possibility of fake block by this equation [36]:

$$\theta = Z \frac{q}{p} \qquad (12)$$

$$P_z = 1 - \sum_{k=0}^{z} \frac{\theta^k e^{-\theta}}{k!} * \left(1 - \frac{q}{p}\right)^{z-k} \qquad (13)$$

Where :

p is honest nodes probability to engender following block.

q is attacker nodes probability to engender following block.

$\mathbf{P_z}$ is attacker probability to draw near the main chain from z blocks behind.

P_z value is calculated for q attacker probability and p honest probability; Fig 13 shows the statistical results.

We conclude the value of z exponentially comes down through the above chart. Blockchain network contains a huge amount of nodes and calculating power, therefore for the attacker to engender a block and change totally nodes records in blockchain, the attacker wants more calculating power than the all blockchain network, which is almost impossible. These results coincide with those provided in reference [36], in which it is proven that it is impossible for an attacker to generate a fraudulent block and alter all records, due to the overwhelming number of nodes and computational power in a blockchain network.

5. CONCLUSION AND FUTURE WORK

The proposed methodology begins with the mathematical foundation of big data-blockchain mapping. Then the proposed big data blockchain infrastructure is proposed. The infrastructure allows many operations such as separating and storing data, querying the separated stored data, block validation, and gossip protocol. A prototype system using the python programming language to confirm that the suggested concept of isolating and storing data is effectively applicative to big data management systems on blockchain technologies. Finally, forged block attack is scrutinized.

This work focused on enhancing the security, scalability, and efficiency of big data management using blockchain, without extending into other blockchain applications such as cryptocurrencies or smart contracts. Future work may consider incorporating other features of blockchain.

REFERENCES:

[1] Jin D, Hannon C, Li Z, et al. Smart street lighting system: A platform for innovative smart city applications and a new frontier for cybersecurity[J]. The Electricity Journal, 2016, 29(10): 28-35.

[2] Boyd D, Crawford K. Six provocations for big data, A decade in internet time: Symposium on the dynamics of the internet and society. Oxford: Oxford Internet Institute, 2011, 21.

[3] Tsakalakis, Niko, S. Stallabourdillon, and K. O'Hara. "What's in a name: the conflicting views of pseudonymisation under eIDAS and the General Data Protection Regulation." European Journal of Psychotraumatology3.2(2016):163-167.

[4] Swan M. Blockchain: Blueprint for a new economy. " O'Reilly Media, Inc.", 2015.

[5] Peters G W, Panayi E. Understanding Modern Banking Ledgers through Blockchain Technologies: Future of Transaction Processing and Smart Contracts on the Internet of Money, Banking Beyond Banks and Money. Springer International Publishing, 2016: 239-278.

[6] Brooke Auxier, Monica Anderson. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021

[7] Moreno, Julio, Manuel A. Serrano, and Eduardo Fernández-Medina. "Main issues in big data security." Future Internet 8.3 (2016): 44.

[8] Swan, Melanie. Blockchain: Blueprint for a new economy. " O'Reilly Media, Inc.", 2015.

[9] Peters, Gareth W., and Efstathios Panayi. "Understanding modern banking ledgers through blockchain technologies: Future of transaction processing and smart contracts on

the internet of money." Banking beyond banks and money. Springer, Cham, 2016. 239-278.

[10] Zheng, Zibin, et al. "Blockchain challenges and opportunities: A survey." International Journal of Web and Grid Services 14.4 (2018): 352-375.

[11] W. Viriyasitavat and D. Hoonsopon, “Blockchain characteristics and consensus in modern business processes,” Journal of Industrial Information Integration, vol. 13, pp. 32–39, Mar. 2019.

[12] L. Da Xu and W. Viriyasitavat, “Application of blockchain in collaborative internet-of-things services,” IEEE Transactions on Computational Social Systems, vol. 6, no. 6, pp. 1295–1305, 2019.

[13] S. Shalini and H. Santhi, “A survey on various attacks in bitcoin and cryptocurrency,” in International Conference on Communication and Signal Processing (ICCSP), 2019, pp. 0220–0224.

[14] J. Parkin, “The senatorial governance of bitcoin: making (de) centralized money,” Economy and Society, vol. 48, no. 4, pp. 463–487, 2019.

[15] F. Casino, T. K. Dasaklis, and C. Patsakis, “A systematic literature review of blockchain-based applications: current status, classification, and open issues,” Telematics and Informatics, vol. 36, pp. 55–81, Mar. 2019.

[16] J. Zhang, S. Zhong, T. Wang, H.-C. Chao, and J. Wang, “Blockchain-based systems and applications: A survey,” Journal of Internet Technology, vol. 21, no. 1, pp. 1–14, Jan. 2020.

[17] Deepa, Natarajan, et al. "A survey on blockchain for big data: approaches, opportunities, and future directions." arXiv preprint arXiv:2009.00858 (2020).

[18] H. Li and D. Han, “EduRSS: a blockchain-based educational records secure storage and sharing scheme,” IEEE Access, vol. 7, pp. 179 273–179 289, 2019.

[19] Y. Chen, J. Guo, C. Li, and W. Ren, “FaDe: a blockchain-based fair data exchange scheme for big data sharing,” Future Internet, vol. 11, no. 11, p. 225, 2019.

[20] H. Hassani, X. Huang, and E. Silva, “Big-Crypto: big data, blockchain and cryptocurrency,” Big Data and Cognitive Computing, vol. 2, no. 4, p. 34, 2018.

[21] J. Liu and Z. Liu, “A survey on security verification of blockchain smart contracts,” IEEE Access, vol. 7, pp. 77 894–77 904, 2019.

[22] N. Tariq, M. Asim, F. Al-Obeidat, M. Zubair Farooqi, T. Baker, M. Hammoudeh, and I. Ghafir, “The security of big data in fog-enabled IoT applications including blockchain: a survey,” Sensors, vol. 19, no. 8, p. 1788, Apr. 2019.

[23] C. G. Akcora, M. F. Dixon, Y. R. Gel, and M. Kantarcioglu, “Blockchain data analytics,” Journal of IEEE Intelligent Informatics, p. 4, 2018.

[24] J. Yang, Z. Lu, and J. Wu, “Smart-toy-edge-computing-oriented data exchange based on blockchain,” Journal of Systems Architecture, vol. 87, pp. 36–48, Jun. 2018.

[25] V. Gramoli and M. Staples, “Blockchain standard: Can we reach consensus?” IEEE Communications Standards Magazine, vol. 2, no. 3, pp. 16–21, Sep. 2018.

[26] F. Hofmann, S. Wurster, E. Ron, and M. Böhmecke-Schwafert, “The immutability concept of blockchains and benefits of early standardization,” in 2017 ITU Kaleidoscope: Challenges for a Data-Driven Society (ITU K). IEEE, 2017, pp. 1–8.

[27] M. Feng, J. Zheng, J. Ren, A. Hussain, X. Li, Y. Xi, and Q. Liu, “Big data analytics and mining for effective visualization and trends forecasting of crime data,” IEEE Access, vol. 7, pp. 106 111–106 123, 2019.

[28] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, “Big data-driven optimization for mobile networks toward 5G,” IEEE network, vol. 30, no. 1, pp. 44–51, Jan.-Feb. 2016.

[29] L. Tan, N. Shi, C. Yang, and K. Yu, “A blockchain-based access control framework for cyber-physical-social system big data,” IEEE Access, vol. 8, pp. 77 215–77 226, 2020.

[30] T. McConaghy, R. Marques, A. Müller, D. De Jonghe, T. McConaghy, G. McMullen, R. Henderson, S. Bellemare, and A. Granzotto, “BigchainDB: a scalable blockchain database,” white paper, BigChainDB, 2016.

[31] C. Esposito, A. De Santis, G. Tortora, H. Chang, and K. R. Choo, “Blockchain: A panacea for healthcare cloud-based data security and privacy?” IEEE Cloud Computing, vol. 5, no. 1, pp. 31–37, Jan 2018.

[32] S. Jangirala, A. K. Das, and A. V. Vasilakos, “Designing secure lightweight blockchain-enabled RFID-based authentication protocol for supply chains in 5G mobile edge computing environment,” IEEE Transactions on Industrial Informatics, vol. 16, no. 11, pp. 7081–7093, 2020.

[33] G. S. Aujla, M. Singh, A. Bose, N. Kumar, G. Han, and R. Buyya, “BlockSDN: blockchain-as-a-service for software defined networking in smart city applications,” IEEE Network, vol. 34, no. 2, pp. 83–91, Mar.-Apr. 2020.

[34] Chen, Jian, Zhihan Lv, and Houbing Song. "Design of personnel big data management system based on blockchain." Future Generation Computer Systems 101 (2019): 1122-1129.

[35] Saldamli, Gokay, et al. "Improved gossip protocol for blockchain applications." Cluster Computing (2022): 1-12.

[36] Yue, Li, et al. "Big data model of security sharing based on blockchain." 2017 3rd International Conference on Big Data Computing and Communications (BIGCOM). IEEE, 2017.

[37] El-Dosuky, Mohamed A., and Gamal H. Eladl. "SPAINChain: security, privacy, and ambient intelligence in negotiation between IoT and blockchain." World Conference on Information Systems and Technologies. Springer, Cham, 2019.

[38] Ogbuke, Nnamdi Johnson, et al. "Big data supply chain analytics: ethical, privacy and security challenges posed to business, industries and society." Production Planning & Control 33.2-3 (2022): 123-137.

[39] Nakamoto, Satoshi. "Bitcoin: A peer-to-peer electronic cash system." Decentralized Business Review (2008): 21260.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

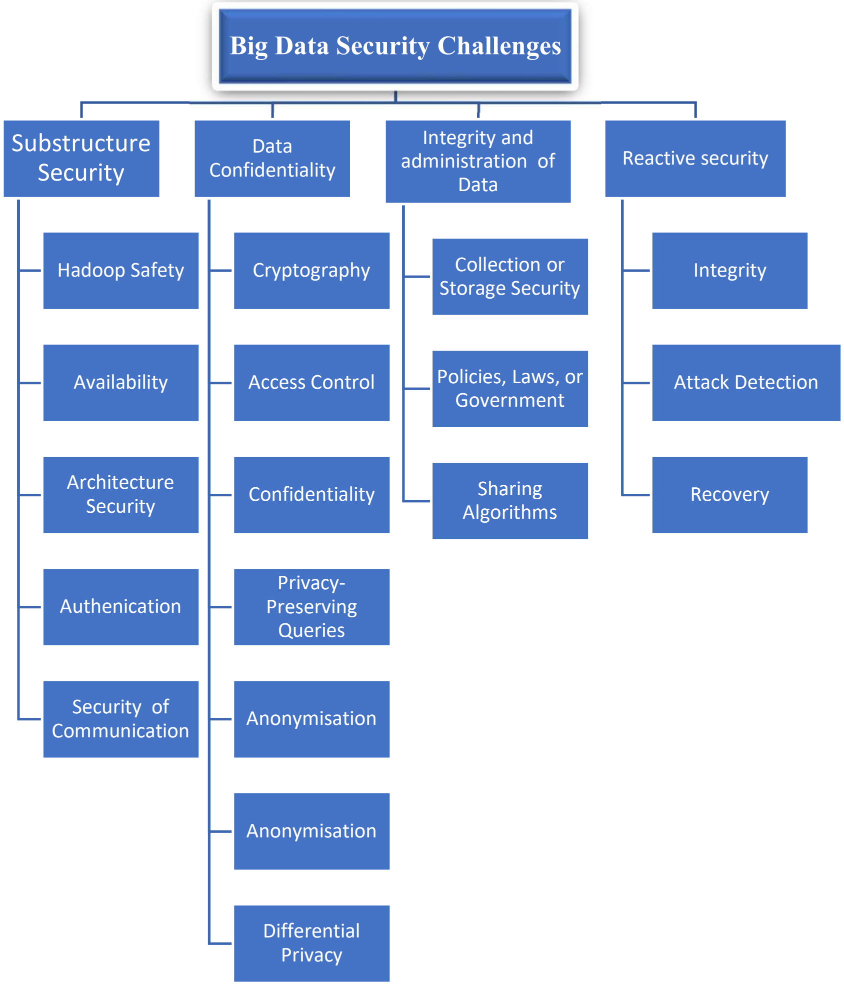

Figure 1: Security Challenges in Big Data [38].

Table 1 Comparison among different blockchain infrastructures [10]

	Consensus determination	**Read Permission**	**Immutability**	**Efficiency**	**Centralized**
Public blockchain	Totally mineworkers	Public	Impossible to alter	Weak	No
Association blockchain	Set of nodes	public or constrained	Possibly altered	Strong	Limited
Private blockchain	Single organization	Public or constrained	Possibly altered	Strong	Yes

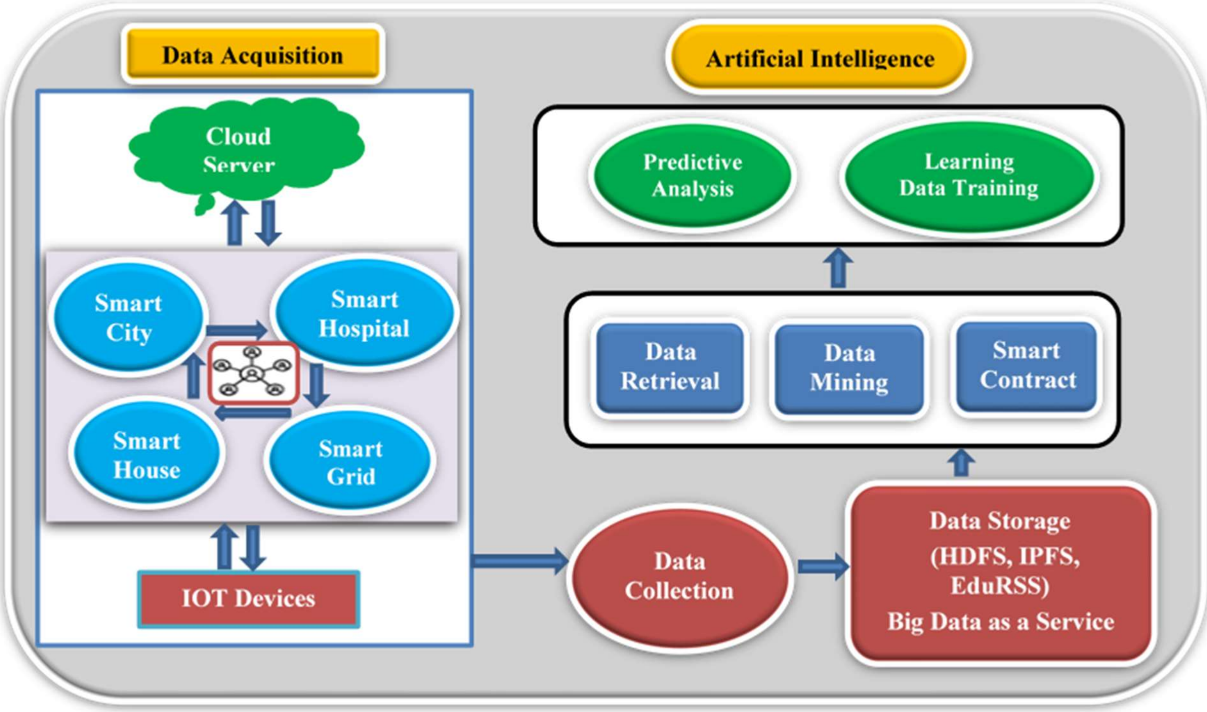

Figure 2 Blockchain services summary in a big data environment [17].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

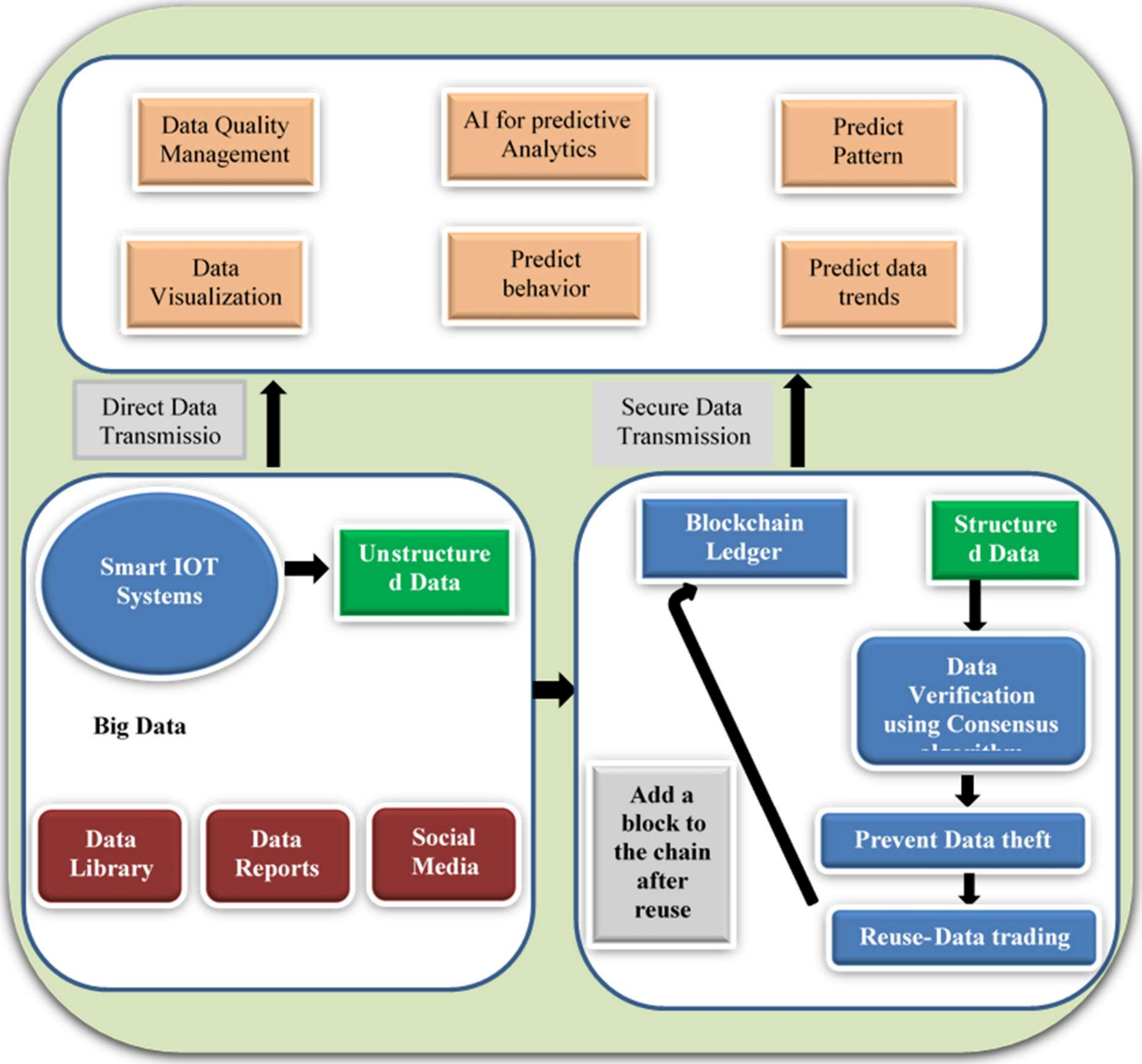

Figure 3. Secured Database Management with blockchain [17].

Table 2: Integration Challenges of blockchain and big data [17]

Reference	Challenges	Application
[20]	Big data Security and Privacy	How to integrate cryptocurrency and big data in decentralized environment.
[21]		Elimination of security holes using smart contracts.
[23]	Big data exchange Security and privacy	How to classify data stored securely in the network of blockchain.
[24]		Blockchain verify massive data interchange through Smart Contract.
[25]	Blockchain standardization	Explain the value of blockchain is immutable

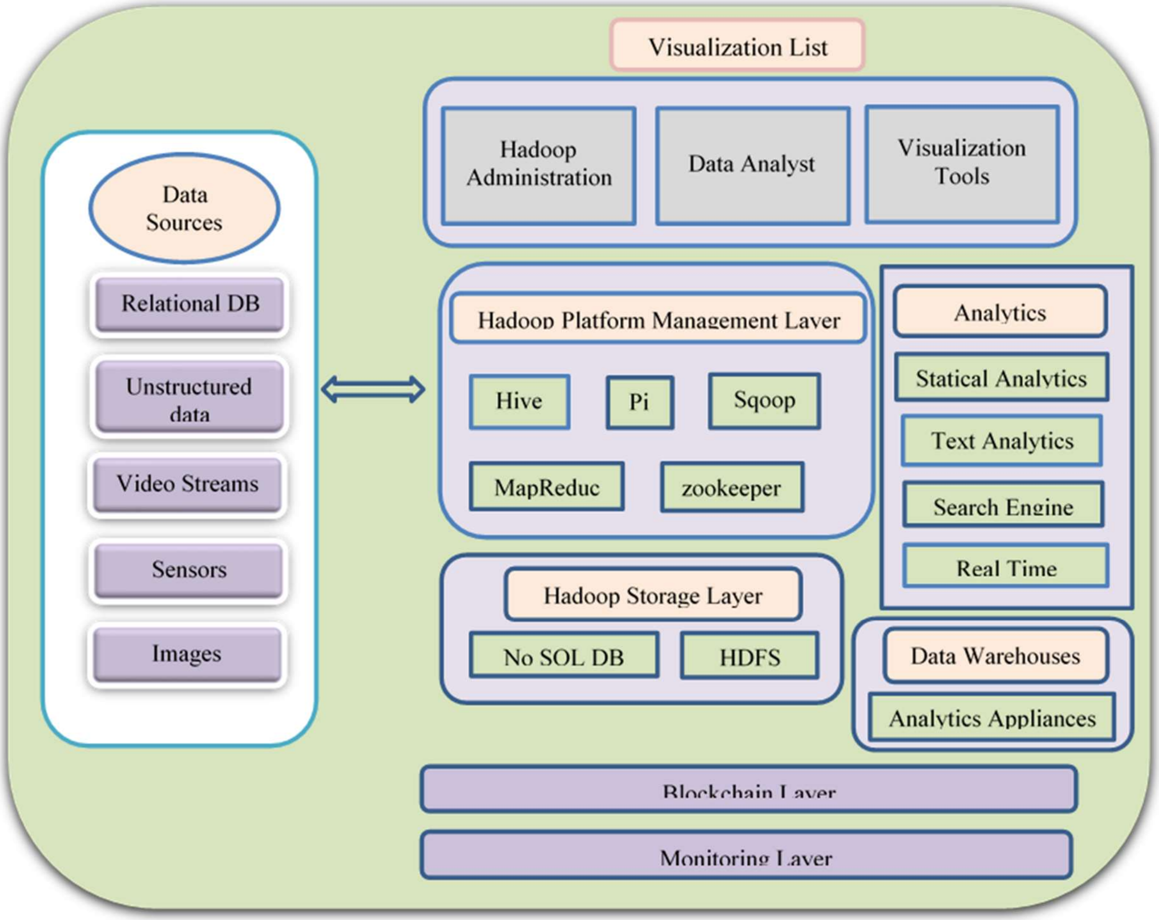

Figure 4: Big data blockchain layer [34]

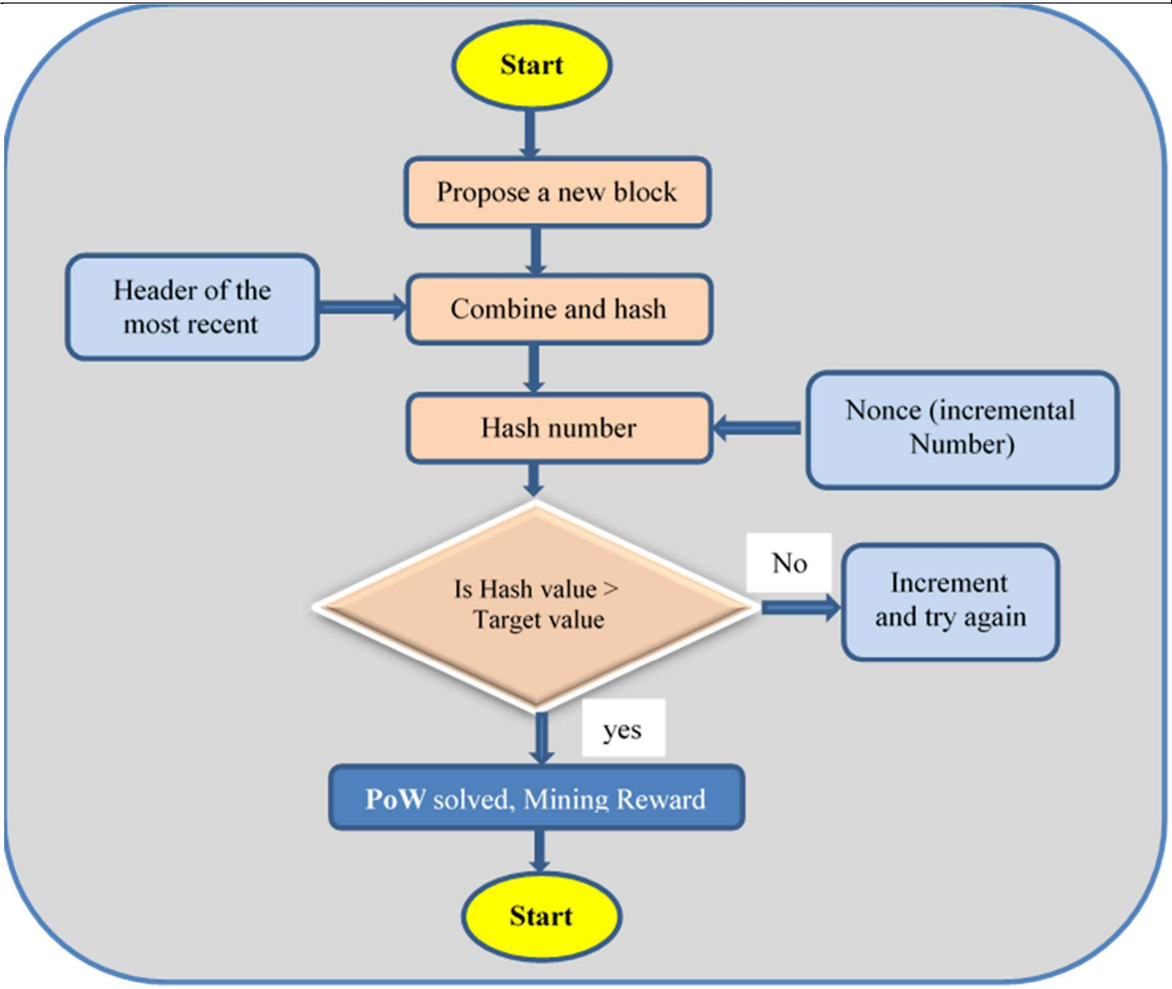
Start
Propose a new block
Header of the most recent
Combine and hash
Hash number
Nonce (incremental Number)
Is Hash value > Target value
No
Increment and try again
yes
PoW solved, Mining Reward
Start

Figure 5: Proof of work

Table 3: Blockchain platforms and relevant properties (1: Minimum advantageous, 2: Fewer advantageous, 3: Extra advantageous, 4: Maximum advantageous) [37].

	Scalability	Consensus	Anonymity	Block size	Smart contract	Security
Bitcoin	1	4	3	1	1	3
Ethereum	4	4	3	4	4	3
Hyperledger	4	4	4	4	4	4
Ripple	4	4	4	4	4	4
Multichain	4	4	4	4	1	4
Eris	3	4	4	4	4	4

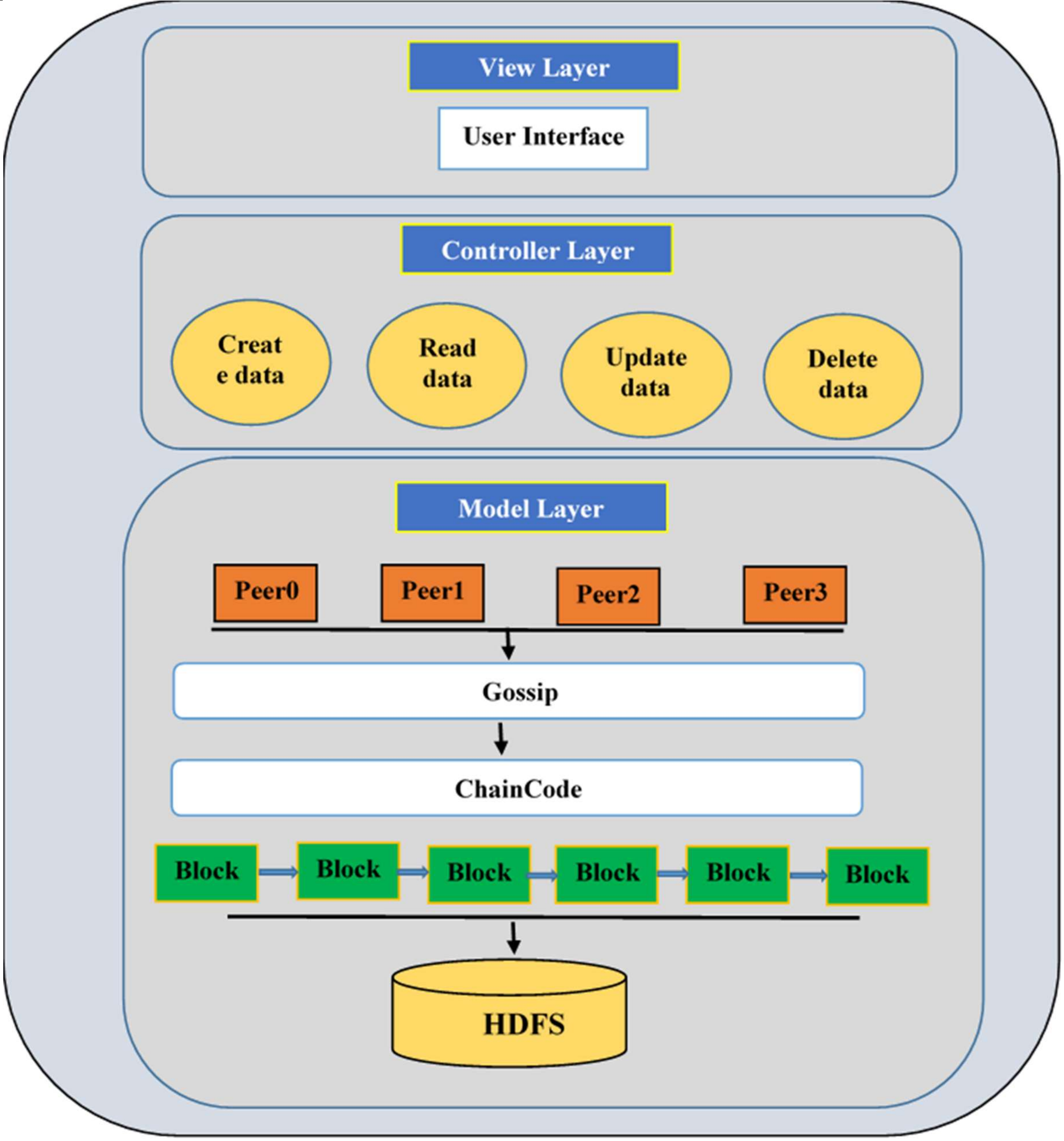

Figure 6: Proposed big data blockchain infrastructure

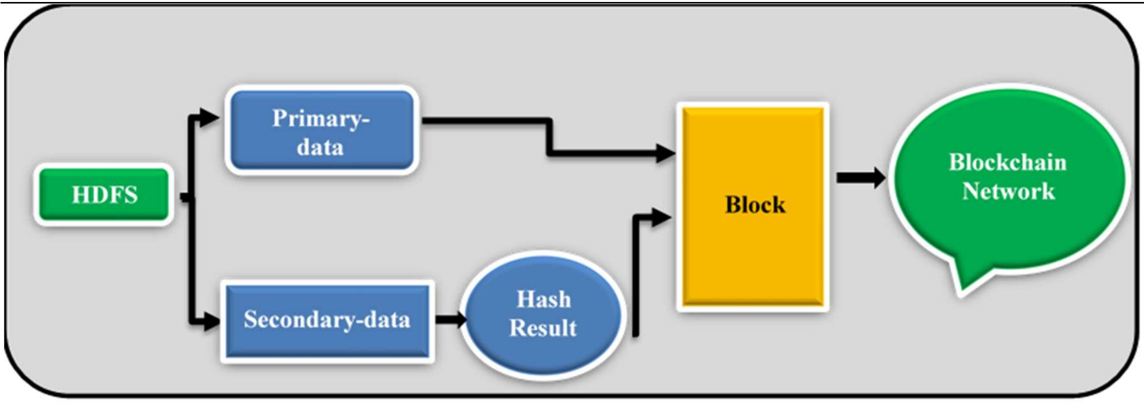

Figure 7: The process of splitting and storing data

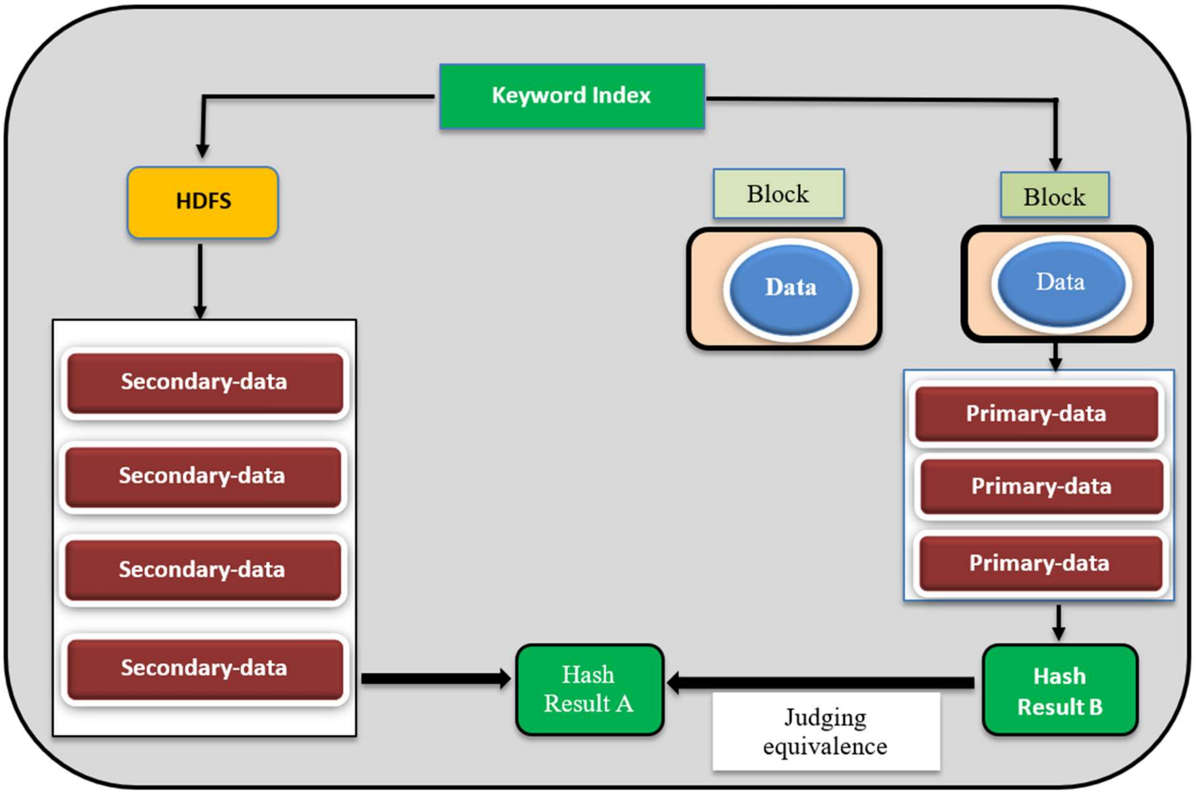

Figure 8: Query a distributed stored data.

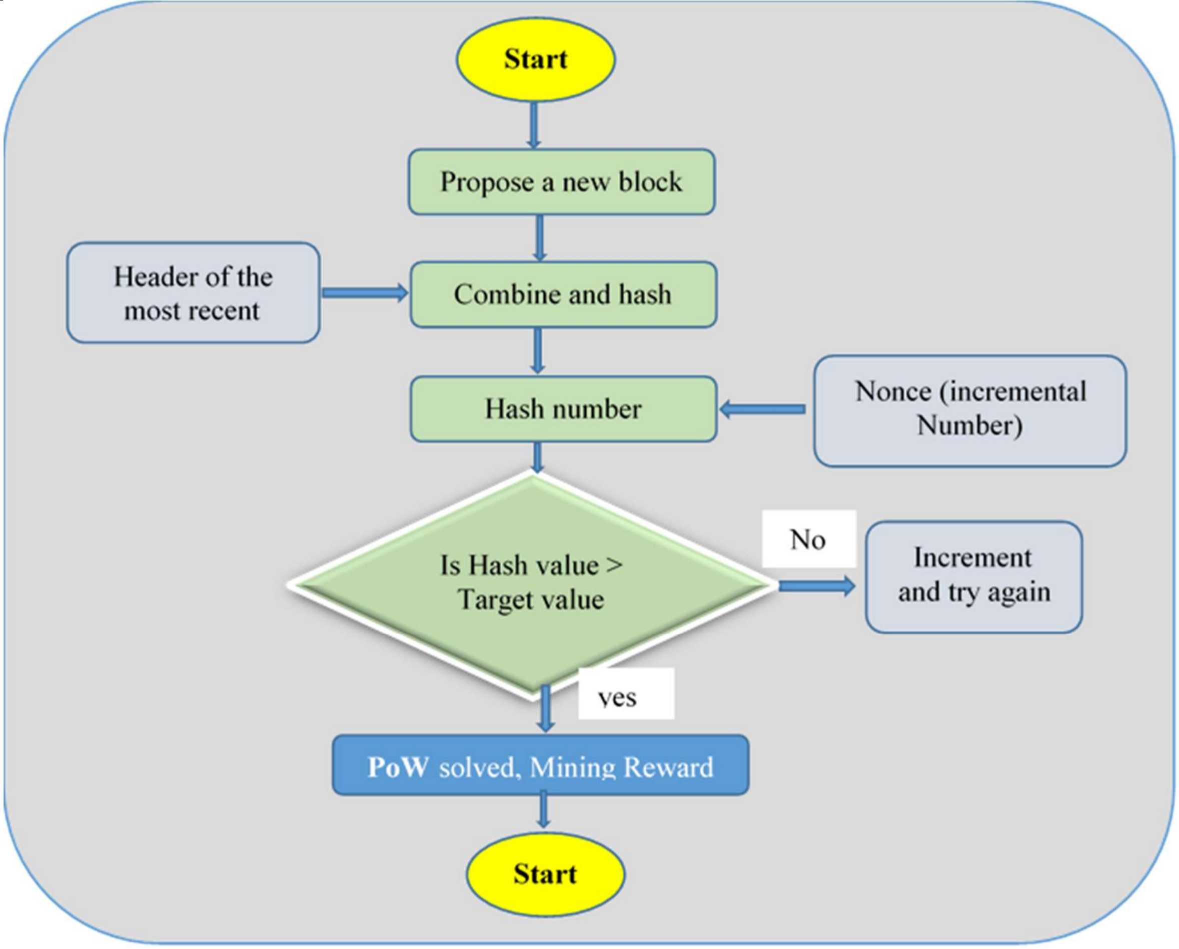
Start
Propose a new block
Header of the most recent
Combine and hash
Hash number
Nonce (incremental Number)
Is Hash value > Target value
No
Increment and try again
yes
PoW solved, Mining Reward
Start

Figure 9: Block Validation

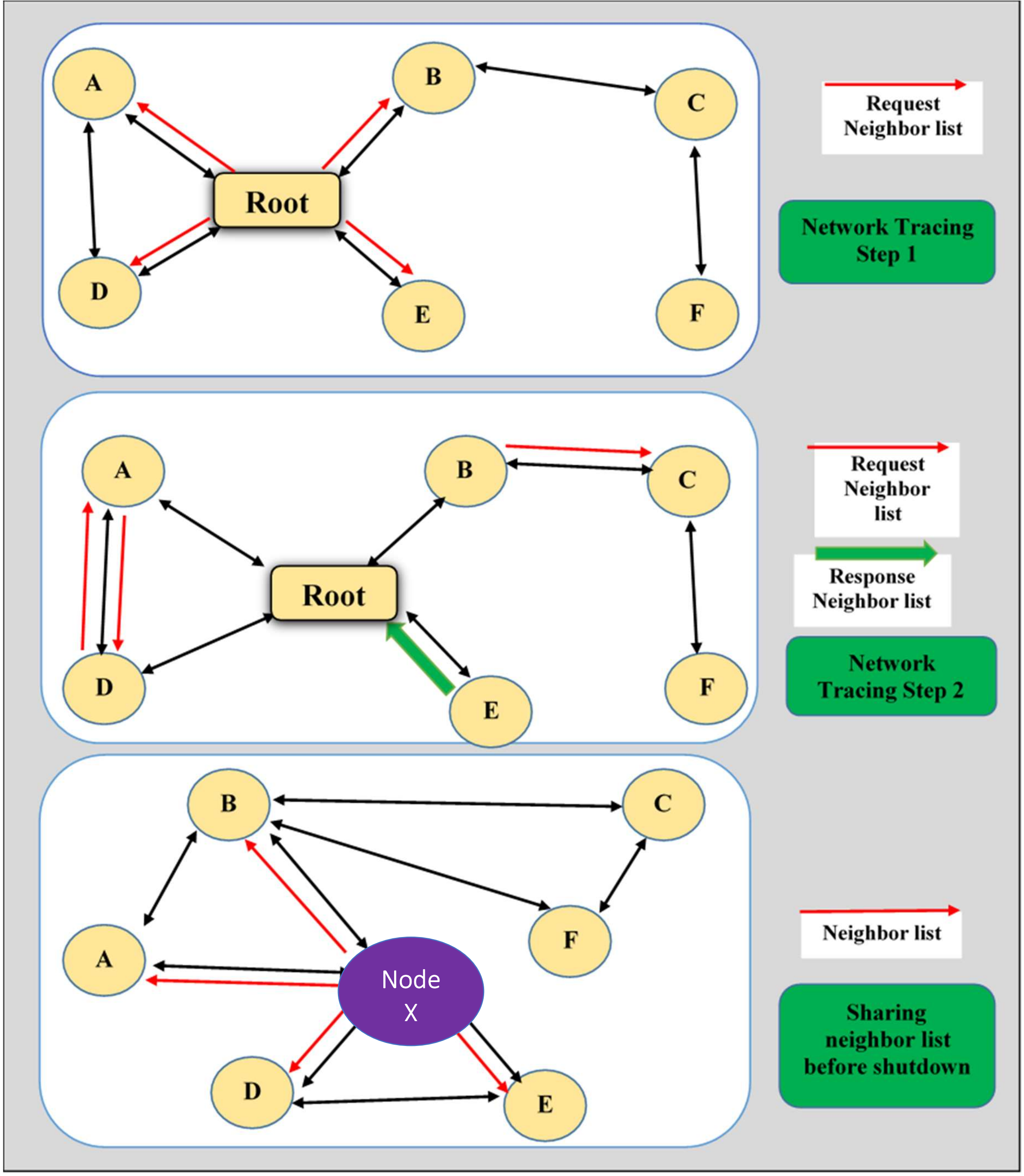

Figure 10: Gossip Protocol Tracing Steps

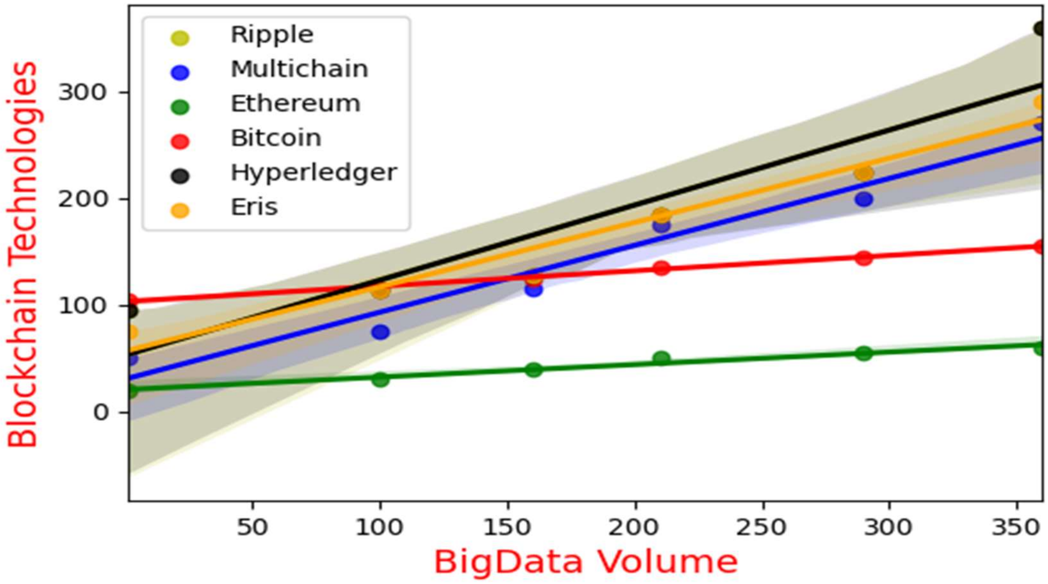

Figure 11: The relationship between the Big Data Volume and the Blockchain Technologies.

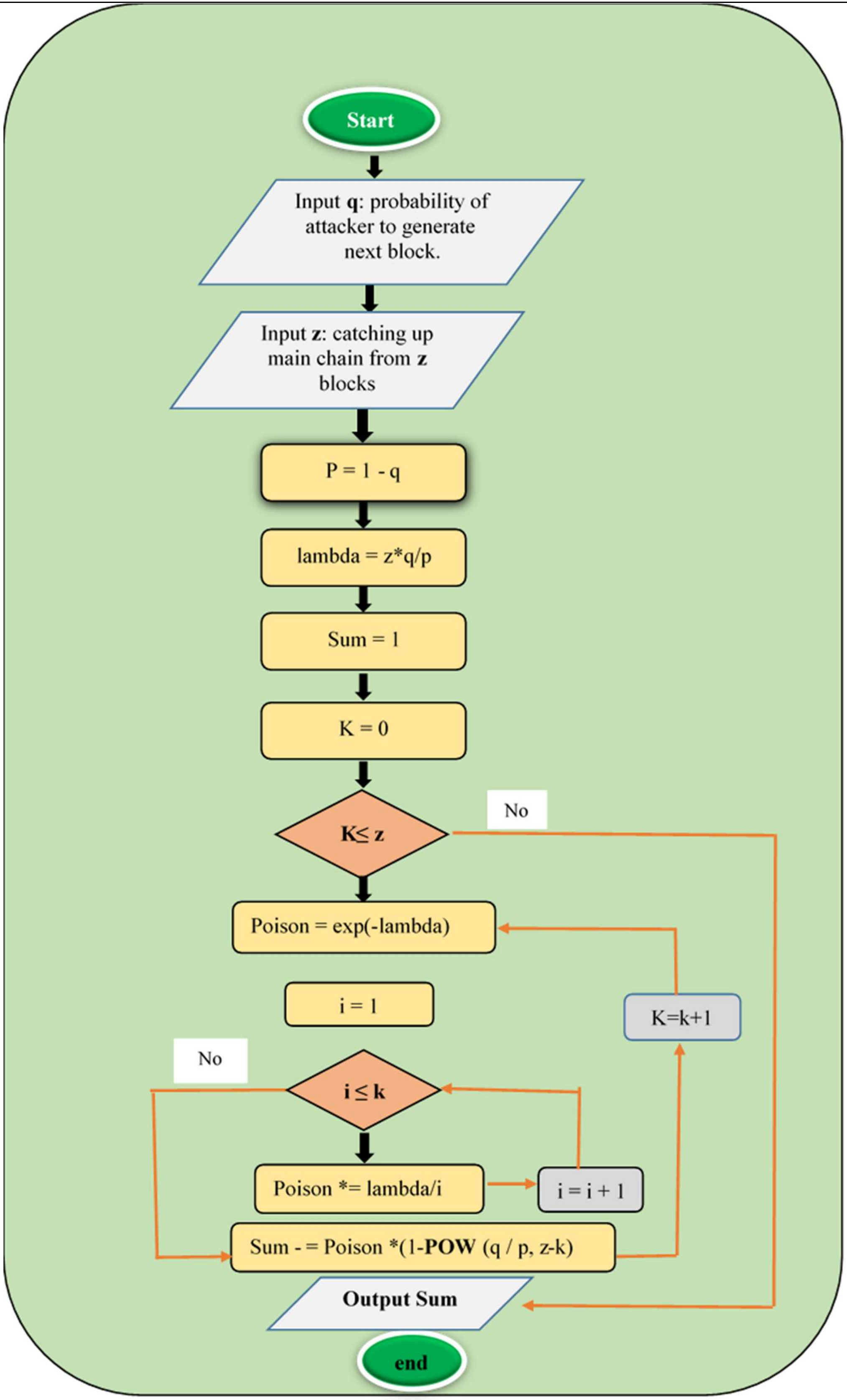
Start
Input q: probability of attacker to generate next block.
Input z: catching up main chain from z blocks
P = 1 - q
lambda = z*q/p
Sum = 1
K = 0
K≤ z
No
Poison = exp(-lambda)
i = 1
K=k+1
No
i ≤ k
Poison *= lambda/i
i = i + 1
Sum - = Poison *(1-POW (q / p, z-k)
Output Sum
end

Figure 12: Forged Block Attack Success Rate

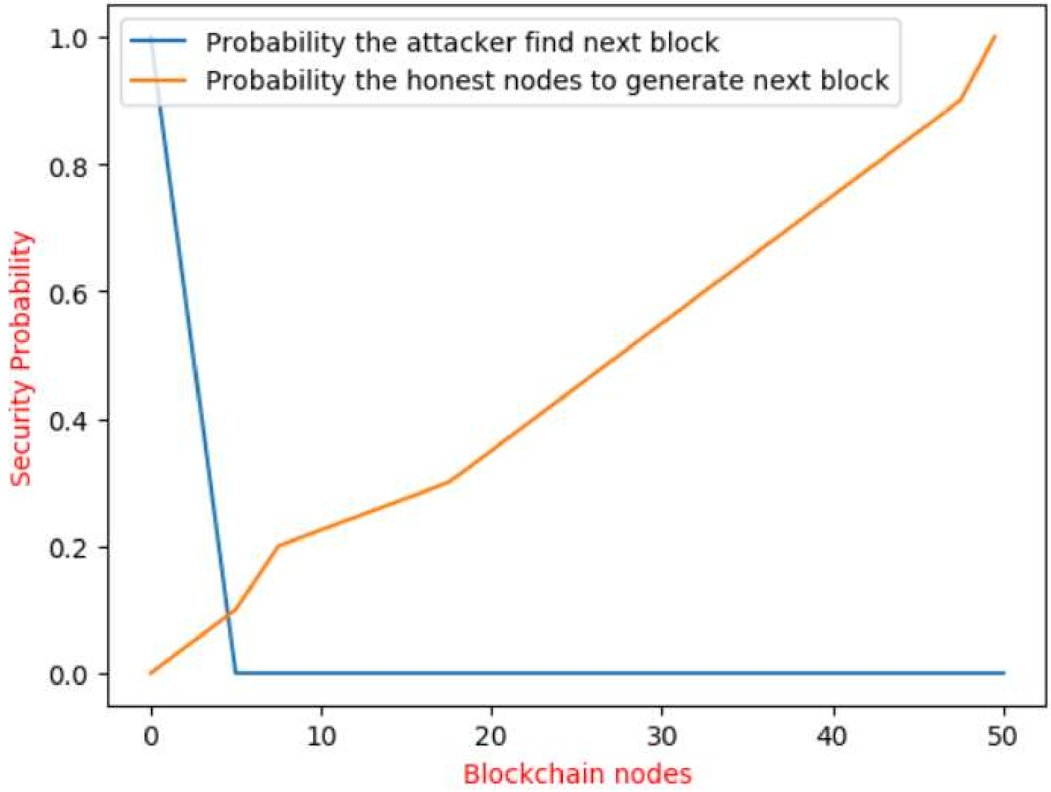

Figure 13. Probability of attacker success.