Designing an Experimental Setup for Data Provenance Tracking using a Public Blockchain: A Case Study using a Water Bottling Plant

O. L. Mokalusi[1], R. B. Kuriakose[2], H. J. Vermaak[3]
Central University of Technology, Bloemfontein, Free State, South Africa

***Abstract*—Data provenance, in an end-to-end supply chain context, refers to the tracking of the origin and history of every raw material, process, packaging and distribution involved in a manufacturing network. The traditional client-server architecture utilised in centralised systems, stores data in a single location, making it vulnerable to single points of failure, data tampering, and unauthorised access. As a result, a lack of data provenance and standardisation for products in a manufacturing supply chain. This leads to a lack of traceability and transparency. Therefore, this article presents a hypothesis that these challenges can be overcome by incorporating data provenance into blockchain-based smart contracts for traceability and transparency. This article is structured such that it firstly discusses data provenance traceability with a focus on the cloud-based storage system architecture domains for data provenance traceability across end-to-end supply chains. Secondly, this article sheds more light on the design of an experimental setup for blockchain-based data provenance traceability in a manufacturing supply chain using a case study of a water bottling plant. Finally, it showcases and discusses the results of the experiments for this purpose.**

***Keywords*—Data provenance; public blockchain; smart contracts; supply chain; smart manufacturing**

I. Introduction

Provenance, also known as lineage [1] or pedigree, refers to the historical record of data and its origins, showing how the data is utilised, processed, represented, stored, and disseminated, by whom and for what purpose [2]. It involves tracking the origins of data and the modifications made to it. It involves tracking the source of information and changes performed on it.

Subsequently, traceability ensures the ability to track the history and origins of product processes [3]. The management of provenance received extensive attention in database systems. The taxonomy of provenance is divided into five divisions as application of provenance, provenance subject, provenance representation, provenance storing and provenance dissemination.

Provenance is crucial in applications such as forensic analysis, as presented by Abiodun et al. [4], as it provides a digital proof for investigation. However, due to the enormous development in data, finding the origin can be a difficult task, leading to emerging studies in various applications such as Scientific Workflow Management Systems (SWfMS) [5], Database Management Systems (DBMS) [6], Hospital Information Systems [7], and supply chain.

This research focuses on the design of an experimental setup public blockchains to track data provenance in a water bottling plant. The smart manufacturing plant produces bottled water on a Make-to Order (MTO) basis. Presently, end-users lack a method to ascertain the water's source, composition, bottling location, and delivery progress. This is as results of limited research on combating the centralisation of manufacturing of bottled water and none when it comes to cost effective public blockchain platform.

This article is structured such that it firstly discusses data provenance traceability with a focus on the cloud-based storage system architecture domains for data provenance traceability across end-to-end supply chains. Secondly, this article sheds more light on the design of an experimental setup for blockchain-based data provenance traceability in a manufacturing supply chain using a case study of a water bottling plant. Finally, showcases and discusses the results of the experiments for this purpose.

II. Cloud-Based Storage Data Provenance Traceability Across End-to-End Supply Chains

The traditional client-server architecture utilised in centralised systems stores data in a single location, making it vulnerable to single points of failure, data tampering, and unauthorised access. As a result, a lack of data provenance and standardisation for products in a manufacturing supply chain can occur, leading to a lack of traceability and transparency.

The problem of ensuring data provenance traceability across end-to-end supply chains for traceability and transparency is a critical issue. It becomes even more challenging when end-users do not have access to centralised storage, making it difficult for them to verify critical data provenance. This was evident during the 2017 listeriosis outbreak in South Africa [8] led to 216 fatalities because contaminated meat products could not be traced back to their origin in time to effect a product recall. Data provenance traceability problems can be classified based on the system architecture as centralised or decentralised, by Kamble et al. [9].

Zafar et al. [10], proposes a general overview of provenance and expanded the technologies of traditional supply chains into sub-domains. These are technologies utilised to log product data provenance associated with product composition and production process data parameters in the supply chain for traceability. The technologies are classified as centralised, including Wireless Sensor Network (WSN), Internet of Things (IoT),

cloud-based storage, and decentralised based storage, such as blockchain.

The establishment of an end-to-end traceability framework, especially in food-sensitive sectors [11] necessitates, a standardised approach handled by the Electronic Product Code Information Service (EPCIS) developed by GS1 [12]. The GS1 global traceability standard sets a minimum set of requirements for traceability in business processes, ensuring an end-to-end traceability framework irrespective of the underlying technology.

The services of GS1 global traceability standard are split into three categories as identity, capture and share. GS1's capture standards are sub-divided into two categories as follows Barcodes and Electronic Product Code (EPC)/Radio Frequency Identification (RFID). One of the major drawbacks of barcodes and RFIDs is that they are linked to a local database. This means that end-users without access to the local database are unable to verify product provenance.

The problem of ensuring data provenance traceability across end-to-end supply chains for traceability and transparency is a critical issue. It becomes even more challenging when end-users do not have access to centralised storage, making it difficult for them to verify critical data provenance. To address these challenges, a decentralised distributed storage is needed, such as a blockchain-based smart contract, which can offer solutions to centralisation by minimising risks with ensuring provenance, traceability, immutability and trust.

III. Blockchain-Based Storage Data Provenance Traceability in a Manufacturing Supply Chain

This section describes blockchain technology and discusses the different types of blockchain platforms.

A. Blockchain Technology

The concept of integrating blockchain technology was initially presented in the *Bitcoin* whitepaper authored by Satoshi Nakamoto in 2008 [12]. The aim of blockchain platforms is to establish a decentralised distributed ledger for time-stamped transactions among various computers in a peer-to-peer network [13], which eliminates the necessity for cloud-based storage that is vulnerable to single point of failure.

Blockchain technology is a write-only decentralised digital ledger, meaning that transactions or data are trackable and irreversible and can only be added but not edited or removed [14]. Every transaction is stored on the blockchain and grouped together in blocks. From a database perspective, the blockchain can be viewed as a blockchain-structured database, where data is packaged into blocks and connected utilising a chain structure.

Blockchain platform block transactions are initiated from a wallet like *MetaMask* [15]. The first block in the chain is known as the genesis block, which has no parent block. These blocks are connected utilising cryptographic principles, providing a secure and tamper-proof ledger for storing data provenance. The newly generated block has the SHA256 algorithm applied to it. The block of the blockchain platform comprises a header and body. The header contains important fields including the timestamp, the hash code of the previous blockchain platform block, transaction hash, and the root of a Merkle tree root [16] (see Fig. 1).

To construct a Merkle tree, the dataset $Data_1$, $Data_2$, $Data_3$ and $Data_4$ are subjected to cryptographic hashing utilising function, such as SHA-256. As depicted in Fig. 1 each parent node in a Merkle tree derives its hash value from its children's nodes $Data_n$. The resulting hashes of node $Hash_1$ and $Hash_2$ are concatenated to create a new parent node $Hash_{[1,2]}$. Similarly, the resulting hashes of nodes *Hash3* and Hash4 are concatenated to create a parent node $Hash_{[3,4]}$. Finally, the resulting hashes of nodes $Hash_{[1,2]}$ and $Hash_{[3,4]}$ are concatenated to create the root hash value node $Hash_{[1,4]}$, which is linked to every value in the tree.

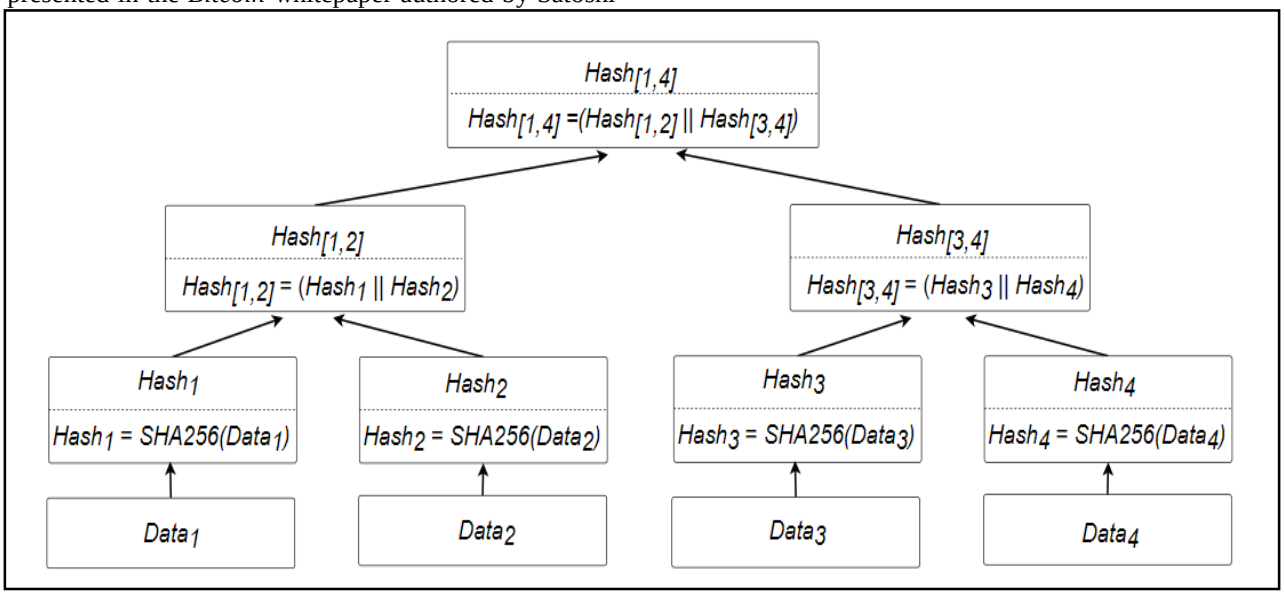

Fig. 1. An illustration of a "balanced" Merkle tree [17].

B. Types of Blockchain Platforms

The three types of blockchain platforms are as follows;

1) Public blockchain: These are permission less [18] network platforms that are accessible to all internet users to interact and view the history of transactions. There is no central authority [19] and all peers can take part in the consensus mechanism to improve data security, integrity, and trust. Public blockchain platforms include and are not limited to *Bitcoin*, *Ethereum*, *Tron* and *Polygon*. However, the efficiency of permissionless public blockchain platforms is low, which affects adoption cost and is their main drawback [20].

2) Private blockchain: These are permissioned [18] network platforms that are centrally managed by a single organization [21]. Private blockchain platforms include and are not limited to *Multichain* [22]. However, since permissioned [18] private blockchain platforms are centralised [23], the decentralised concept is therefore compromised [20] and this is their main drawback.

3) Consortium blockchain: These network platforms have permissioned or permissionless access authority for end-users and can be public or private. They are considered partially decentralised [24] since only organisations parts of the network platform have access permission authority. Consortium blockchain network platforms include and not limited to Hyperledger, Corda and Quorum. However, because permissioned consortium blockchain platforms are centralised, the decentralised concept is compromised, and this is their main drawback.

Table I compares public, private, and consortium blockchain platform types [25] [26] such as Ethereum (legacy), Ethereum 2, Polygon, and Tron.

TABLE I. THE COMPARISON BETWEEN BLOCKCHAIN TYPES BY TORKY AND HASSANEIN 2020

Blockchain type	Public	Private	Consortium
Centralisation	No	Yes	Partial
Access authority	Permissionless	Permissioned	Permissioned
Efficiency	Low	High	High
Trust	All peers	Single organisation	Selected peers

IV. METHODOLOGY

The aim of this article is to design an experimental setup for data provenance tracking using a public blockchain. In this research study, the water bottling plant located at the Bloemfontein campus of the Central University of Technology (CUT) [27] was chosen as the case study (see Fig. 2).

The proposed solution for data provenance collection involves the utilisation of Near Field Communication (NFC) tags. NFC antennas are mounted at designated points A, B, C, and D. The NFC antennas are mounted at the start and end of each Smart Manufacturing Unit (SMU), designated for water filling, capping, and packaging activities. Each SMU activity triggers a Critical Tracking Event (CTE), which is then incorporated into a blockchain-based smart contract to ensure data security, immutability, and traceability.

In order to facilitate the design process, the experimental setup of the water bottling plant (see Fig. 2) is split into three subsystems. It is important to note that certain functions of the water bottling plant, such as capping, labeling, and packaging have been omitted from the experimental setup for convenience (see Fig. 3).

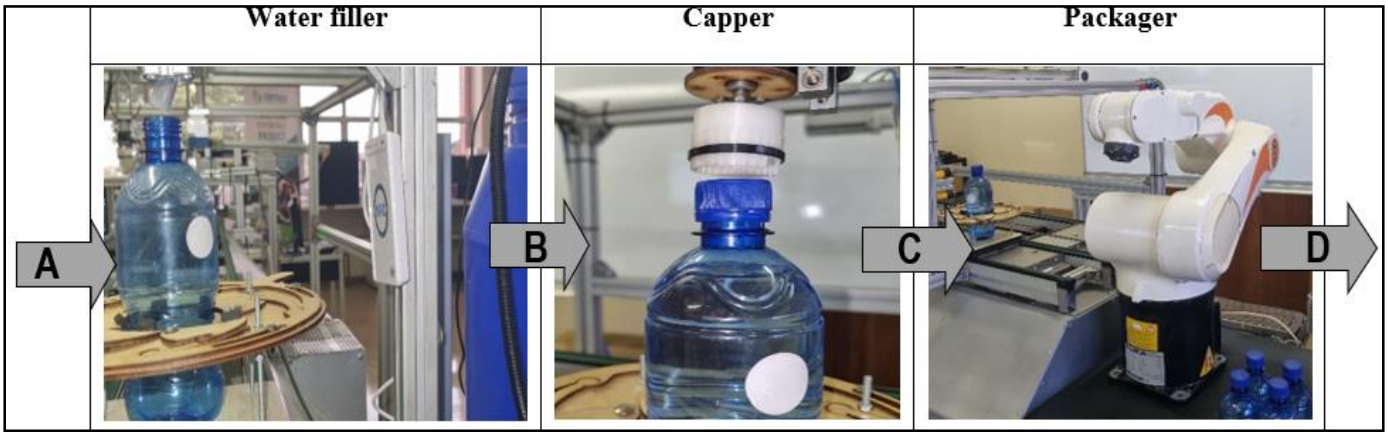

Fig. 2. The schematic outline of a water bottling plant at CUT.

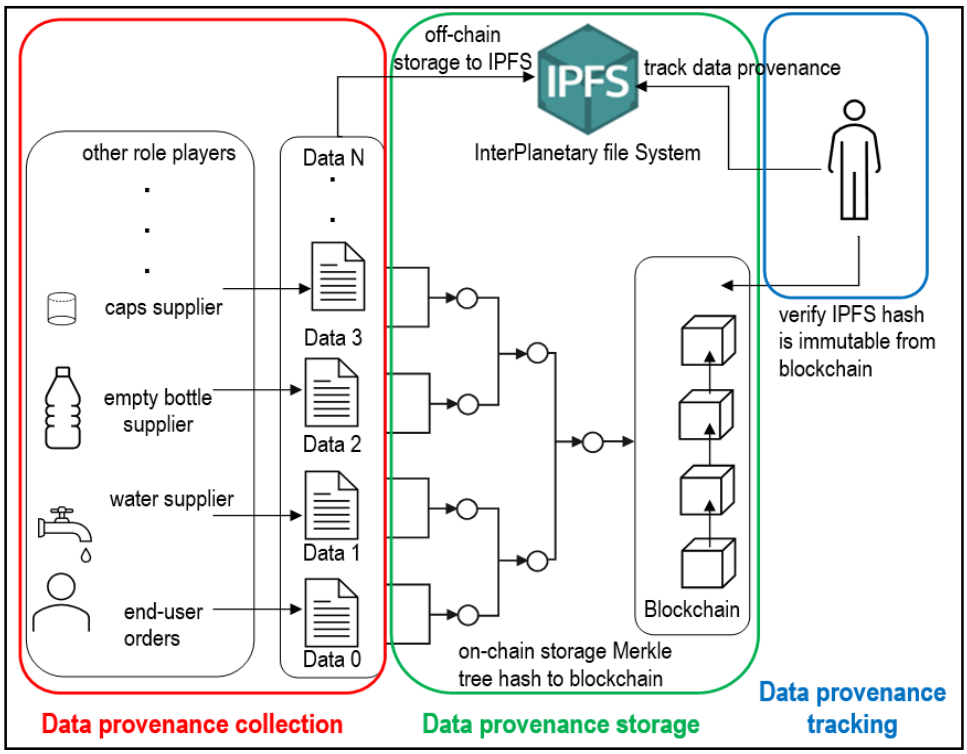

Fig. 3. The block diagram of the experimental setup.

V. Results

This section describes the results of the methodology discussed in Section IV. Firstly, the type of collected data provenance is described. Next, the analysis focuses on the results of data provenance stored. Finally, the results of the data provenance tracking are shared.

A. Data Provenance Collection

The data collected is written on the Near Field Communication (NFC) tag in a JavaScript Object Notation (JSON) markup format and shared as object of $Data_n$ (Data N), where N is an index set starting from zero. The NTAG21x [28] series are utilised, which are compatible with most NFC-enabled devices. Each role player data is collated through an NFC tag with its Unique Identifier (UID).

These input sources include the end-user's new order, water supplier and empty bottle supplier. The end-user provides order requirements, followed by the water sourced from a specific supplier, and the empty bottles and caps from other suppliers. The Fig. 4 describes a JSON markup format array *new_order*, the first entry object of Data0. If additional role players need to be added, provisions can be made.

B. Data Provenance Storage

A study that was done into the factors that influences the selection of a blockchain platform [26] resulted in this input. The efficiency of permissionless public blockchain platforms is low, which can increase adoption costs. Therefore, the Inter-Planetary File System (IPFS) is utilised to store large-sized data provenance by generating content-addressed hash with a fixed length. The role player sources are stored in the IPFS as node $Hash_0$ (see Fig. 5).

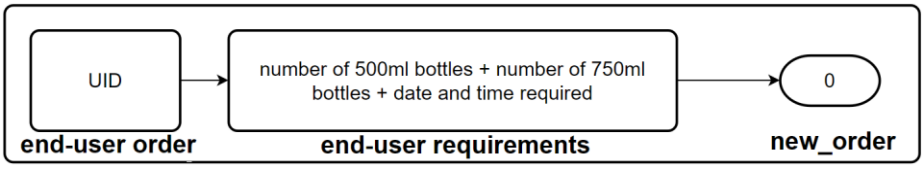

Fig. 4. End-user's new order data provenance collection.

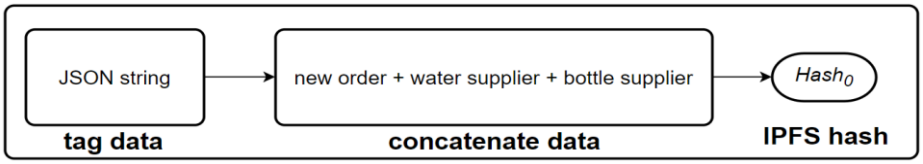

Fig. 5. The role player's sources storing in the IPFS.

The concept of Merkle tree Directed Acyclic Graph (DAG) is utilised in IPFS. LineageChain, the study in [29] is a fine-grained data provenance traceability solution for blockchains platforms to achieve efficient tracking and querying of semantic information. The authors captured verifiable semantic information during the execution of blockchain-based smart contracts and stored it in a Merkle tree, which was converted into a Directed Acyclic Graph (DAG).

Therefore, the principles of the IPFS Merkle tree DAG are implemented with Smart Manufacturing Unit (SMU), tasked with filling the water bottles. IPFS Merkle hash tree DAG node is created by utilising Python IPFS Application Programming Interface (API) which accepts data provenance as a JSON markup format that is stored with the function *IPFS DAG PUT* to obtain a unique content-addressed transaction hash.

Once the blockchain-based smart contract is deployed, its generated content-addressed transaction hash is written on the NFC tag as a Uniform Resource Locator (URL) link. This data provenance was stored off-chain in the IPFS Merkle tree DAG and its root hash incorporated into the blockchain-based smart contract.

C. Data Provenance Tracking

The ability to track and query the IPFS Merkle tree DAG, facilitates end-to-end supply chain tracking, enabling verification of product history and origin to ensure the integrity of data provenance. The end-user can read data provenance by tapping a Near Field Communication (NFC) tag attached to the bottled water, which contains a content-addressed transaction hash Uniform Resource Locator (URL) link. A URL link directs the end-user to the deployed blockchain-based smart contract transaction hash on the Polygon public blockchain platform.

This is to retrieve the root InterPlanetary File System (IPFS) Merkle tree Directed Acyclic Graph (DAG) node which represents a Critical Tracking Event (CTE) for water filling which is immutable. A top-down data provenance tracking approach is achieved by the ability to traverse the IPFS Merkle tree DAG with node $Hash_{[1,3]}$ (see Fig. 6).

Firstly, the links from the parent node $Hash_{[1,3]}$ are then verified, with two children links. The one pointing to node $Hash_{[1,2]}$ of the measured water pH level in the tank. The other as a node $Hash_3$ of the counted water bottles. Follow the link to node $Hash_{[1,2]}$ and retrieve its content utilising its IPFS Merkle tree DAG hash, which has two children links pointing to node $Hash_1$ and $Hash_2$. The content corresponding to the new_order is retrieved by following the link to node $Hash_1$. Similarly, the content corresponding to the water_supply is retrieved by following the link to node $Hash_2$. Once the contents of node $Hash_1$ and node $Hash_2$ have been retrieved, navigate back to node $Hash_{[1,3]}$ and follow its links to node $Hash_3$, which includes the content corresponding to the empty bottle_supply.

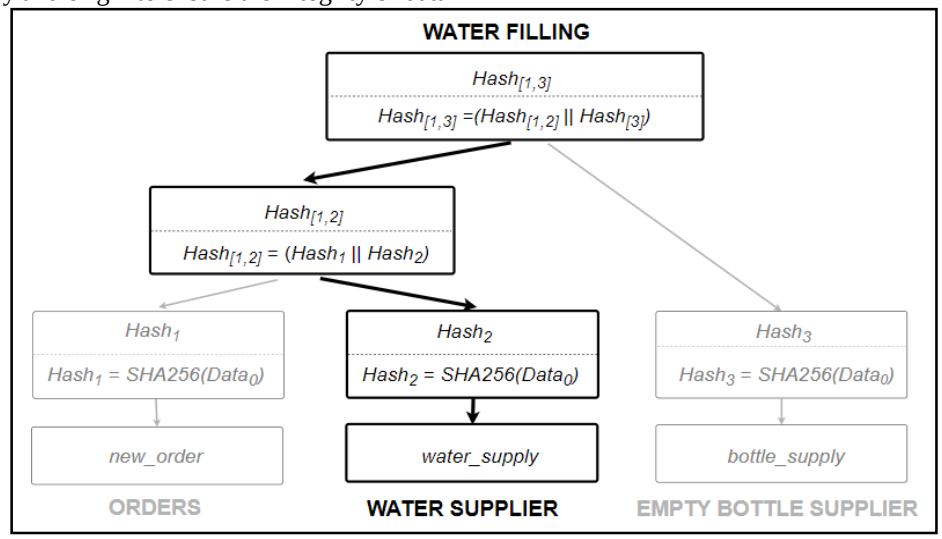

Fig. 6. IPFS Merkle tree DAG inversion method for data provenance tracking.

VI. Discussion and Conclusion

This article investigated the broader problem of data provenance traceability with a focus on the cloud-based storage system architecture domains for data provenance traceability across end-to-end supply chains. It then identified the challenges described and established the limitations that this research study aims to fill. There is a limited or lack research on combating the centralisation of manufacturing of bottled water and none when it comes to cost effective public blockchain platform. The proposed experimental setup ensures that data provenance pertaining to the end-user order, raw materials used, and timestamp at each stage of production is incorporated into a blockchain smart contract to make it immutable and traceable. This will allow the end-user to "read" data provenance, through a tag attached on the bottled water for traceability and transparency. The results of this study can also be seen as an addition to the knowledge base of the broader studies on data provenance traceability which can reduce the size of a recall from millions to just a few hundred units, highlighting the importance of this research study.

References

[1] Kuftinova, N.G., Ostroukh, A. V, Maksimychev, O.I., Pronin, C.B., Yadav, A.K.: Implementation of the Data Fabric Architecture as a Sustainable Development of Industrial Platform Technologies in Road

Transport Systems. In: 2023 Systems of Signals Generating and Processing in the Field of on Board Communications. pp. 1–5 (2023).

[2] Khan, S.N., Loukil, F., Ghedira-Guegan, C., Elhadj Benkhelifa, •, Anoud Bani-Hani, •: Blockchain smart contracts: Applications, challenges, and future trends. (2021). https://doi.org/10.1007/s12083-021-01127-0.

[3] Treiblmaier, H., Garaus, M.: Using blockchain to signal quality in the food supply chain: The impact on consumer purchase intentions and the moderating effect of brand familiarity. Int J Inf Manage. 68, 102514 (2023). https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2022.102514.

[4] Isaac Abiodun, O., Alawida, M., Esther Omolara, A., Alabdulatif, A.: Data provenance for cloud forensic investigations, security, challenges, solutions and future perspectives: A survey. Journal of King Saud University - Computer and Information Sciences. (2022). https://doi.org/10.1016/J.JKSUCI.2022.10.018.

[5] da Cruz, S.M.S., Campos, M.L.M., Mattoso, M.: Towards a taxonomy of provenance in Scientific Workflow Management Systems. SERVICES 2009 - 5th 2009 World Congress on Services. 259–266 (2009). https://doi.org/10.1109/SERVICES-I.2009.18.

[6] Buneman, P., Davidson, S.B.: Data provenance-the foundation of data quality. (2010).

[7] Vázquez-Salceda, J., Álvarez, S., Kifor, T., Varga, L.Z., Miles, S., Moreau, L., Willmott, S.: EU PROVENANCE Project: An Open Provenance Architecture for Distributed Applications. Agent Technology and e-Health. 45–63 (2008). https://doi.org/10.1007/978-3-7643-8547-7_4.

[8] Tchatchouang, C.D.K., Fri, J., De Santi, M., Brandi, G., Schiavano, G.F., Amagliani, G., Ateba, C.N.: Listeriosis outbreak in south africa: A comparative analysis with previously reported cases worldwide, /pmc/articles/PMC7023107/, (2020).

[9] Kamble, S.S., Gunasekaran, A., Sharma, R.: Modeling the blockchain enabled traceability in agriculture supply chain. Int J Inf Manage. 52, (2020). https://doi.org/10.1016/J.IJINFOMGT.2019.05.023.

[10] Zafar, F., Khan, A., Suhail, S., Ahmed, I., Hameed, K., Khan, H.M., Jabeen, F., Anjum, A.: Trustworthy data: A survey, taxonomy and future trends of secure provenance schemes. Journal of Network and Computer Applications. 94, 50–68 (2017). https://doi.org/10.1016/J.JNCA.2017.06.003.

[11] Behnke, K., Janssen, M.F.W.H.A.: Boundary conditions for traceability in food supply chains using blockchain technology. Int J Inf Manage. 52, 101969 (2020). https://doi.org/10.1016/j.ijinfomgt.2019.05.025.

[12] Dasaklis, T.K., Voutsinas, T.G., Tsoulfas, G.T., Casino, F.: A Systematic Literature Review of Blockchain-Enabled Supply Chain Traceability Implementations. Sustainability 2022, Vol. 14, Page 2439. 14, 2439 (2022). https://doi.org/10.3390/SU14042439.

[13] Srikanth, M., Mohan, R.N.V.J., Naik, M.C.: Blockchain based Crop Farming Application Using Peer-to-Peer. xidian journal. 16, 168–175 (2022).

[14] Cui, P., Dixon, J., Guin, U., Dimase, D.: A Blockchain-Based Framework for Supply Chain Provenance. IEEE Access. 7, 157113–157125 (2019). https://doi.org/10.1109/ACCESS.2019.2949951.

[15] Duran, R.S., Balbon, C.B., Balmes, J.Z.E., Castro, P.B.G., Lopez, U.G., Talosig, K.A., Noriega, Ma.E.A., Tubola, O.D.: EASY E-VOTE: An Ethereum-based Voting System using IPFS and MetaMask for Student Government Election. In: 2023 International Conference on Information Technology (ICIT). pp. 815–820 (2023).

[16] Merkle, R.C.: A digital signature based on a conventional encryption function. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 293 LNCS, 369–378 (1988). https://doi.org/10.1007/3-540-48184-2_32/COVER.

[17] Huang, H., Lin, J., Zheng, B., Zheng, Z., Bian, J.: When Blockchain Meets Distributed File Systems: An Overview, Challenges, and Open Issues. IEEE Access. 8, 50574–50586 (2020). https://doi.org/10.1109/ACCESS.2020.2979881.

[18] Risso, L.A., Ganga, G.M.D., Godinho Filho, M., de Santa-Eulalia, L.A., Chikhi, T., Mosconi, E.: Present and future perspectives of blockchain in supply chain management: a review of reviews and research agenda. Comput Ind Eng. 179, 109195 (2023). https://doi.org/https://doi.org/10.1016/j.cie.2023.109195.

[19] Jabbar, S., Lloyd, H., Hammoudeh, M., Adebisi, B., Raza, U.: Blockchain-enabled supply chain: analysis, challenges, and future directions. Multimed Syst. 27, 787–806 (2021). https://doi.org/10.1007/S00530-020-00687-0/FIGURES/9.

[20] Rouhani, S., Deters, R.: Performance analysis of ethereum transactions in private blockchain. Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS. 2017-November, 70–74 (2018). https://doi.org/10.1109/ICSESS.2017.8342866.

[21] Zheng, Z., Xie, S., Dai, H., Chen, X., Wang, H.: An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends. Proceedings - 2017 IEEE 6th International Congress on Big Data, BigData Congress 2017. 557–564 (2017). https://doi.org/10.1109/BIGDATACONGRESS.2017.85.

[22] Hegde, P., Maddikunta, P.K.R.: Amalgamation of Blockchain with resource-constrained IoT devices for healthcare applications – State of art, challenges and future directions. International Journal of Cognitive Computing in Engineering. 4, 220–239 (2023). https://doi.org/https://doi.org/10.1016/j.ijcce.2023.06.002.

[23] Ameyaw, P.D., de Vries, W.T.: Transparency of Land Administration and the Role of Blockchain Technology, a Four-Dimensional Framework Analysis from the Ghanaian Land Perspective. Land 2020, Vol. 9, Page 491. 9, 491 (2020). https://doi.org/10.3390/LAND9120491.

[24] Dey, S., Saha, S., Singh, A.K., McDonald-Maier, K.: FoodSQRBlock: Digitizing Food Production and the Supply Chain with Blockchain and QR Code in the Cloud. Sustainability 2021, Vol. 13, Page 3486. 13, 3486 (2021). https://doi.org/10.3390/SU13063486.

[25] Torky, M., Hassanein, A.E.: Integrating blockchain and the internet of things in precision agriculture: Analysis, opportunities, and challenges. Comput Electron Agric. 178, (2020). https://doi.org/10.1016/J.COMPAG.2020.105476.

[26] Mokalusi, O.L., Kuriakose, R.B., Vermaak, H.J.: Factors Influencing the Selection of a Blockchain Platform for Incorporating Data Provenance into Smart Contracts. 517–525 (2023). https://doi.org/10.1007/978-981-19-2394-4_47.

[27] Kuriakose, R.B., Vermaak, H.J.: Designing a Simulink model for a mixed model stochastic assembly line : A case study using a water bottling plant. Journal of Discrete Mathematical Sciences and Cryptography. 23, 329–336 (2020). https://doi.org/10.1080/09720529.2020.1741184.

[28] Anthony, Lee, M.C., Pearl, R.R., Edbert, I.S., Suhartono, D.: Developing an anti-counterfeit system using blockchain technology. Procedia Comput Sci. 216, 86–95 (2023). https://doi.org/10.1016/J.PROCS.2022.12.114.

[29] Ruan, P., Dinh, T.T.A., Lin, Q., Zhang, M., Chen, G., Ooi, B.C.: LineageChain: a fine-grained, secure and efficient data provenance system for blockchains. The VLDB Journal 2021 30:1. 30, 3–24 (2021). https://doi.org/10.1007/S00778-020-00646-1.