Zeppelin Server: A Novel Approach to Data Sharing and Digital Identity Management

Ishan Joshi, Ameya Mahabaleshwarkar, Sudheendra Katikar, Sujay Mahadik

Abstract— Personal data has gained paramount importance with the emergence of data driven technologies that can reveal trends and patterns which help businesses optimize their operations for better quality results and create personalized experiences for customers. In a scenario where every major business is chasing personal data, users are rendered vulnerable to privacy invasion and information misuse. There have been many cases in which unauthorized third parties have, intentionally or unintentionally, accessed personal information of individuals for their own benefit. Consequently, innovation in secured data sharing systems has become the need of the hour. Conventional digital identity management and data sharing systems have several vulnerabilities which can allow crucial information to be exploited. To address this issue, the novel design presented in this paper utilizes cutting edge technologies to form a more robust architecture that ensures secure data sharing without compromising the performance of business applications. Blockchain technology, due to its inherent decentralized, resilient and scalable nature becomes a viable technology upon which such a system canbe based. The system presented in this paper makes use of the Ethereum blockchain to allow user consent to be introduced in the data request and sharing process. The user authorization is placed centrally in the data transaction surrounded by two blockchains that handle the data request and data access functionality. InterPlanetary File System has been used to introduce a mechanism for sharing the authorized files in an immutable and secured manner. The system presented thus attempts to innovate the data sharing process itself, such that the data can serve its purpose without actually being completely revealed.

Index Terms— Blockchain, Data Sharing & Security, InterPlanetary File System, Ethereum, Merkel DAG, Distributed Hash Tables, Smart Contracts, GDPR, Peer to Peer File Sharing

1 INTRODUCTION

Upcoming technologies like Big Data and Data Analytics require good quality, genuine data to yield good results [1]. Their ability to provide competitive advantage by increasing efficiency and profits has seen number of businesses chase after user data to fuel these technologies [2]. Personal data has also become central to several applications that require identity verification and authentication like Know YourCustomer (KYC) [3]. This increase in demand for data has resulted in several incidents where data was shared in an unauthorized manner, without the awareness of its owners [3],[4]. As a result, several laws and regulations are now being implemented to protect users' privacy such as Global Data Protection Regulation (GDPR) by the European Union (EU) [4]. In spite of the several privacy and data security concerns, the potential of these data driven technologies in revolutionizing critical sectors like disaster management, medicine and healthcare cannot be undermined [6], [7], [8].

Thus, data sharing with high level security and privacy towards the Data Owner has become an area of increasing concern so as to allow data to be used in genuine applications without violating Data Owners' privacy at the same time. This paper presents how Blockchain can be used as a robust and secure platform, upon which a mechanism for secured data sharing and privacy protection can be built. Our system model aims at balancing the two seemingly contradictory notions of availability of useful data and privacy protection by emphasizing on controlled access to information under users' consent. Our system utilizes blockchains to facilitate the data transaction process of requesting data and granting or denying access to the data by its owner, while also providing secured data sharing mechanisms to prevent malicious access to data.

The following sections aim to increase the readers' understanding of blockchain and InterPlanetary File System and demonstrate how our system model uses these two disruptive technologies for secured data sharing and privacy protection**.**

2 LITERATURE SURVEY

2.1 Blockchain

Blockchain is an emerging peer-to-peer technology that enables decentralized transactions while maintaining an immutable record or ledger of the transactions. Blockchain is nothing but a glorified linked list, in which every block stores a transaction as rows. Each block stores three attributes, the timestamp, transaction details and a hash for the current transaction along with the hash for the previous transaction [9]. A sample diagram of a Blockchain is shown in Fig. 1. The timestamp and the hash of the previous transaction, allows the peers in the network to track the history and verify any transaction. The hash is an encrypted string that gets populated during the transactions, hides the data of the transactions, thus providing security against parties not involved in the transaction from having access to transaction details [10]. Transactions can only be appended to the ledger, whereas modifications and deletions on existing transactions are not permitted [9]. A transaction is validated, by using a consensus algorithm, which is computed by the peers in the network. Only after validation, a block containing the transaction is created, and, appended to the Blockchain. The process of validation is computationally heavy, making it more secure as more resources are allocated to it. This process is often incentivized and the participating peers are rewarded for validating transactions [11].

In traditional transaction systems, data was essentially centralized and maintained by a third-party trusted organization. The structure of blockchain is able to provide a robust, secure and decentralized environment for recording and maintaining data, that eliminates the dependence on a centralized third-party [9]. Thus, blockchains can play a pivotal role in secure data sharing.

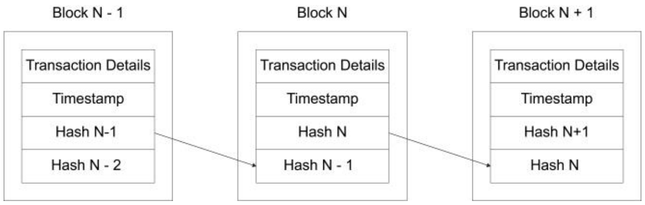


Fig. 1: Structure of Blockchain

2.2 Peer to Peer (P2P) File Sharing

Applications of P2P file sharing such as BitTorrent leverage its users' resources to distribute all types of digital files to its consumers without the need of a central governing model. Client server architecture-based content sharing services often incurs high electricity cost to maintain high speeds of content delivery, to maintain the temperature of the servers among many other factors [12]. Since P2P architecture is self-maintaining, resilient and only needs limited infrastructure and control, it is vastly superior, faster, more secure and robust than the existing client server architecture [13], [14].

2.3 Distributed Hash Table

Distributed Hash Tables are used as a lookup service in distributed and decentralized services [13]. They are used to map identifiers from a common pool of peers or nodes in an overlay network [15]. A DHT is an extension of a simple hash table that saves data in the form of key-value pairs on Node IDs. The Node Id is generated using the nodes' IP address or geographic information. The key is generated using a custom hash function using the data item as a parameter [13]. Most of the existing DHT assumes that its peers are spread over the ID space uniformly [15].

2.4 Interplanetary File System (IPFS)

The Interplanetary File System is a decentralized file system, distributed across multiple peers all over the world forming an interconnected network of nodes. The peer to peer architecture used for content delivery enables IPFS to distribute zettabytes of files with ease and efficiency while managing the network traffic. IPFS is a file system similar to UNIX, and can store files of any format. This gives applications the flexibility to use custom file formats [16]. IPFS employs Distributed Hash Tables (DHTs), which are spread over a network of well-coordinated computers that enable efficient lookup between nodes and provide a robust and scalable architecture. Messages sent from an IPFS node are routed through a peer of nodes, while recording the IP Addresses and node IDs of involved peers in DHTs called Routing Tables [16], [17]. IPFS uses a data structure called Merkle Directed Acyclic Graphs (DAGs) to produce a hash for each file, which can be used to retrieve the file. Another powerful feature of the Merkle DAG structure is that it allows you to build a distributed version control system [16]. The three main advantages of Merkel DAGs that are used by IPFS are [16]:

Content Addressing: While adding any file to the IPFS network, a Merkle DAG is produced by using the contents of the file. At the root of this DAG is a hash, which is used for retrieving the same file.

Tamper Resistance: Since, the file can only be retrieved by its hash, if any malicious user changes the file, the Merkle DAG of this modified file would generate a different hash. This makes the network tamper proof.

Deduplication: The same file would generate the same Merkle DAG, meaning, the same hash. Hence, IPFS can prevent re-uploading the same file to the network, and can store the same file only once. Thus, the aforementioned advantages make IPFS a secure and scalable platform for file transfer. You can get

2.5 Consensus Protocols

Consensus Protocols are the backbone of any blockchain application. Such protocols are used to provide authenticity, non-repudiation and integrity to the blockchain network, by utilizing a decentralized peer-to-peer network for verification of transactions before adding a block to the public ledger [18, 19]. The bitcoin blockchain uses the concept of Proof of Work (PoW) to help decide validate the transactions occurring and also helps in avoiding the forking problem in blockchain [18, 19]. Some other types of Consensus Protocols are Delegated Proof of Stake (DPoS), Proof of Activity (PoA) - an amalgamation of PoW and PoS [18].

2.6 Smart Contracts

A smart contract is "a digital contract that is written in source code and executed by computers, which integrates the tamper-proof mechanism of blockchain" [17], [20]. Smart contracts have transformed the blockchain scenario from a financial transaction protocol to an all-purpose utility. They are pieces of software, not contracts in the legal sense, that extend blockchains utility from maintaining a ledger of financial transactions to automatically implementing conditions of multi-party agreements. Smart contracts are executed by a computer network that uses consensus protocols to agree upon the series of actions resulting from the contracts content [21]. The high-level programming languages used for writing smart contracts are mainly Solidity, Serpent and Low-level Lisp-like Language (LLL) [20].

2.7 Ethereum

Ethereum is an open source blockchain capable of running decentralized applications. These decentralized applications are written as smart contracts using a programming language called Solidity [17], [21]. The deployed contracts are hosted by each node in the network. Ethereum nodes use the Proof of Stake consensus protocol as opposed to the Proof of Work protocol used by Bitcoin. The Proof of Stake algorithm was chosen because it adds more security to the network, as well as bringing down the resources required to compute the validity of the transaction [21], [22]. The Ethereum blockchain uses Ether as its cryptocurrency for rewarding the transaction validators or miners. Each operation on the blockchain costs some overhead, which is known as 'gas', whose price is fixed based on the current value of Ether [8], [10]. While reading data from the blockchain is free, the cost of writing data and transferring Ether drastically increases with scalability of the application [23]. Due to the structured validation protocols, Ethereum is suitable for creating a secure data sharing application.

2,8 Interplanetary File System (IPFS)

General Data Protection Regulation came into effect on 25th May 2018 across the European Union. The regulation was introduced to highlight the importance of consent regarding privacy and sharing of data [24]. Prior to GDPR, it was not mandatory to report to data owners that their data is being collected or shared. As an effect, incidences, like sharing of personal user data by Facebook with Cambridge Analytica without permission, went unchecked. GDPR, in order to prevent such instances, makes it a requirement to provide notices to the users highlighting the use of their personal data and the effect that it may have on their privacy, so that the data owner can make an informed decision whether or not to consent the storage and sharing of their data [25]. GDPR focuses on informing data owners about how their data is stored, how to rectify incorrect data and how to delete data that was acquired without permission, thus enabling data owners to exercise a higher degree of control over personal data. GDPR gives an active role to data owners in the data storage and sharing process, and mandates organizations, that collect and store data, to proactively take part in data protection legislation and safeguarding the privacy of data owners [24].

2.9 Increasing Privacy Concerns with Data

The development and progress of technologies and algorithms that can generate value from data has led to an increase in the attention received by genuine, high quality and personalized data. However, a lack of clear-cut guidelines and rules that can govern the collection, storage and access to this data has left individuals concerned about their privacy. According to a survey by McAfee, a security technology giant, more than 40% of the worldwide population believes that they do not have adequate control over their private and personal data [26]. Although in the past few years' rules governing data access and control have emerged, simultaneously major incidents of security breaches have also come to light, leaving everyone question the compliance of major organizations, that heavily depend on data-intensive processes to fuel their products and services, with the upcoming data related laws. Last year, Facebook, the world's largest social media network experienced a data breach that made tens of millions of personal data records vulnerable to illegal, and even criminal exploitation [27]. Another incident occurred at Sacramento Bee, where 19.5 million voter records were leaked [28]. A similar breach at Panera Bread was responsible for exposing 37 million customer records and the company was unaware of this attack for 8 months [29]. As seen from these examples, along with the ease with which data can now be productively used, the risks of losing control over sensitive data has also increased and all of this calls for new innovations in the privacy domains.

3 Methodology

3.1 Stakeholders

The data transaction, as facilitated by our system, can be viewed as an interaction between three entities. The three entities have specific roles to play, which must be properly carried out for the transaction to successfully go through. The three stakeholders and their roles can be described as following -

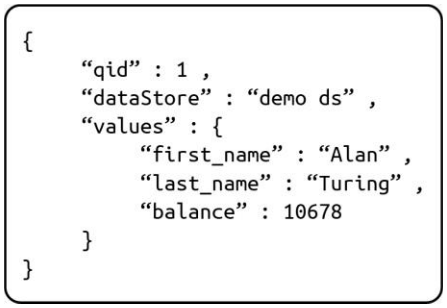
Data Owner: Data owner is the individual whose data is being shared between two organizations for facilitating business services. The Data Owner receives a request specifying who is requesting access to what data. The Data Owner must then authorize or deny the request.

Data Consumer: Data consumer is the institution which is requesting access to the Data Owner's credential for business purposes or verification such as Know Your Customer (KYC).

Data Provider: Data Provider is the organization which is responsible for storage of Data Owner's personal information. Data Provider is sent the authorized request from the Data Owner and must perform the task of making available the data, as per the request, to the Data Consumer.

3.2 Process for Data Sharing

Our proposed system allows for sharing of three types of information. Data can be shared in the form of a 'Yes' or 'No' interpretation, actual credentials or large files. The data sharing request is facilitated in the form of a questionnaire, a copy of which is also shared with the data provider beforehand. The questionnaire allows for a 'Yes' or 'No' response which satisfies consumer requirement while protecting sensitive data. The questionnaire provides easy interpretation for the Data Consumer's request which contains minimal information such as Question ID and querying value. A sample request for checking whether the bank balance of Alan Turing is greater than or equal to 10678 is shown in Fig. 2.



```
{
     “qid” : 1 ,
     “dataStore” : “demo ds” ,
     “values” : {
          “first_name” : “Alan” ,
          “last_name” : “Turing” ,
          “balance” : 10678
     }
}
```

Fig. 2: Sample Request

The response given by the Data Provider for this sample request would be a simple 'Yes' or 'No' depending on the bank balance of Alan Turing. Thus, the business objective of the bank is satisfied without ever revealing the actual bank balance of Alan Turing.

Requesting Data: The data consumer selects a questionnaire based on the specific user credentials that are required and provides the parameters that need to be tested or verified. The system then initiates the process of getting users' consent by passing the corresponding Question ID and querying parameters inside a Hypertext Transfer Protocol (HTTP) POST request which is handled by a Representational State Transfer (REST) Application Programming Interface (API), responsible for converting the Web2.0 request to a Web3.0 request.

The Request Blockchain forwards this request to the concerned Data Owner for authorization. Data Owner has the ability to either authorize or reject the data request. On authorization, the Request Blockchain then contacts the Data Provider with the Question ID and querying parameters.

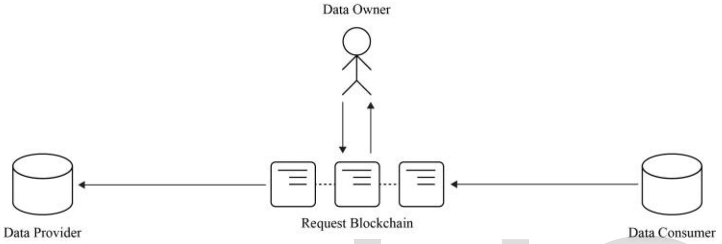


Fig. 3(a): Request and Authorization

Accessing Data: Data Provider after receiving the Question ID and querying parameters from the Request Blockchain, queries its internal database and provides the result to the Response Blockchain. This event is recorded by the Response Blockchain and the data is further shared with the Data Consumer.

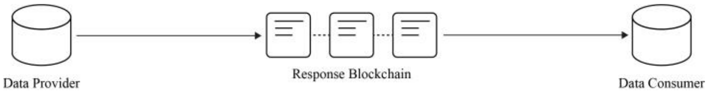


Fig. 3(b): Response and Data Transfer

In case of file sharing, the Data Provider uploads the files on the IPFS and retrieves the content-addressable hash. After the hash is retrieved the Data Provider forwards this information along with the Question ID to the Response Blockchain. The Response Blockchain now emits this data to the Data Consumer, thus completing the data sharing process.

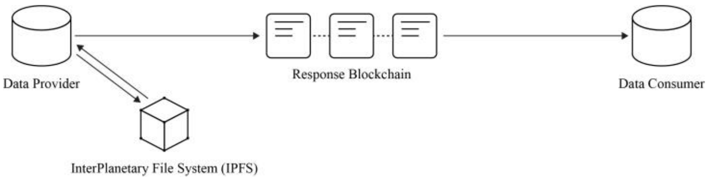


Fig. 3(c): Response and Data Transfer – File Sharing

4 Implementation

The implemented module is a Web 2.0 to Web 3.0, ECMAScript 7 compatible Representational State Transfer (REST) Application Programming Interface (API). The purpose of this API is to enable legacy systems to use the Blockchain based Web 3.0 architecture. The REST API was developed using a node package called Express.js. Along with Express.js, we used the Web3.js module to enable interaction with the Ethereum Blockchain. Interplanetary File System network was accessed using their NodeJS module, by using their read and write methods.

For developing the smart contracts, we used Solidity, an Open Source language, which can run on the Ethereum Blockchain. As a tool manager, we also used the Truffle Framework, which is able to compile the solidity smart contracts, and deploy them on the Ethereum Blockchain. It also offers a module to test the smart contracts before deploying them to the Blockchain network, making it a complete package for Ethereum development.

The test bed consisted of three servers running on different network ports, where each port was assigned to a different stakeholder. This enabled us to incorporate existing technologies and framework namely the React framework, on which the Users' frontend application was developed.

The GET and POST Request were made with the help of Postman, an API Development Software, that handles sending the REST requests and shows their response in a well-designed Graphical User Interface (GUI).

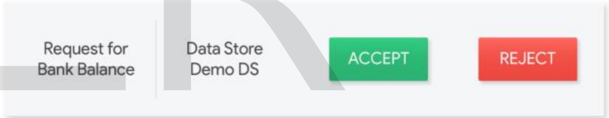


Fig. 4: User Authorization for Data Request(User Interface)

5 Conclusion

This paper has thus presented a system architecture that is able to solve the critical and complex problem of secured data sharing. The system integrates currently used Web2.0 architecture with Blockchain based Web 3.0 architecture, which will allow existing organization to easily adopt the system without having to overhaul their currently functioning infrastructure. With the emerging stringent laws for regulating Data sharing and privacy, our system will help organizations to carry out essential business operations without having to constantly monitor for regulation law violations and violating the Data Owners' privacy. The system also makes sure that the data transactions are always carried out with regard to the Data Owners' consent.

6 Future Work

Future directions of this work would involve work on the encryption used in the system by using the Whisper communication protocol, developed upon Ethereum. Another prospect for expansion of this work is working on an automated questionnaire generation mechanism, which can handle situations in which generation of simple 'Yes' or 'No' questions become complex and tedious, benefitting the Data Consumers' business processes without compromising on the Data Owners' privacy.

REFERENCES

[1] Hand, D. J.: Principles of Data Mining. Drug Safety, 30(7), 621–622 (2007).

[2] John Walker, S.: Big Data: A Revolution That Will Transform How We Live, Work, and Think. International Journal of Advertising, 33(1), 181–183 (2014).

[3] Lootsma, Y.: Blockchain as the Newest Regtech Application – the Opportunity to reduce the Burden of KYC for Financial Institutions. Banking & Financial Services Policy Report 36, 16–21 (2017).

[4] Bhaimia, S.: The General Data Protection Regulation: The Next Generation of EU Data Protection. Legal Information Management, 18(1), 21-28 (2018).

[5] Ouyang, Y., Zhang, J. E., Luo, S. M.: Dynamic data driven application system: Recent development and future perspective. *Ecological Modelling* (2007).

[6] Imran, M., Castillo, C., Lucas, J., Meier, P., Vieweg, S.: AIDR - Artificial Intelligence for Disaster Response. In *WWW'14 Companion* (2014).

[7] Lavecchia, A.: *Machine-learning approaches in drug discovery: methods and applications. Drug Discovery Today, 20(3), 318–331* (2015).

[8] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I.: *Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17.* (2015).

[9] Yli-Huumo, J., Ko, D., Choi, S., Park, S.: Where is Current Research on Blockchain Technology? – A Systematic Review. PLoS ONE, 11(10), e0163477 (2016).

[10] Pierro, M. D.: What is the blockchain? Computing in Science & Engineering, 19 (5), 9295 (2017).

[11] Kondor, D., Pósfai, M., Csabai, I., & Vattay, G. Do the Rich Get Richer? An Empirical Analysis of the Bitcoin Transaction Network. PLoS ONE, 9(2), e86197 (2014).

[12] Pattal, M. M. I., Li, Y., & Zeng, J. (2009). Web 3.0: A Real Personal Web! More Opportunities and More Threats. 2009 Third International Conference on Next Generation Mobile Applications, Services and Technologies.

[13] Nath, K., Dhar, S., & Basishtha, S. (2014). Web 1.0 to Web 3.0 - Evolution of the Web and its various challenges. 2014 International Conference on Reliability Optimization and Information Technology (ICROIT).

[14] J. Blackburn, K. Christensen," A Simulation Study of a New Green BitTorrent," Proc. Green Communications Workshop in conjunction with IEEE ICC'09 (GreenComm09), Dresden, Germany, June 2009.

[15] Landsiedel, O., Gotz, S., & Wehrle, K. (n.d.). Towards Scalable Mobility in Distributed Hash Tables. Sixth IEEE International Conference on Peer-to-Peer

Computing (P2P'06).

[16] Benet, J.: IPFS – Content addressed, version, P2P File system. Technical Report (2014).

[17] Vujičić, D., Jagodić, D., Ranđić, S., Blockchain Technology, Bitcoin, and Ethereum: A Brief Overview, IEEE, 17th International Symposium Infotech-Jahorina (2018).

[18] Zhong, L., Wang, X., & Kihl, M. (2011). Topological model and analysis of the P2P BitTorrent protocol. 2011 9th World Congress on Intelligent Control and Automation.

[19] Han, D., Kim, H., & Jang, J. (2017). Blockchain based smart door lock system. 2017 International Conference on Information and Communication Technology Convergence (ICTC).

[20] Christidis, K., & Devetsikiotis, M. (2016). Blockchains and Smart Contracts for the Internet of Things. IEEE Access, 4, 2292–2303.

[21] Sreehari, P., Nandakishore, M., Krishna, G., Jacob, J., & Shibu, V. S. (2017). Smart will be converting the legal testament into a smart contract. 2017 International Conference on Networks & Advances in Computational Technologies (NetACT).

[22] Khan, N., Lahmadi, A., Francois, J., State, R., Towards a Management Plan for Smart Contracts: Ethereum Case Study, IEEE, IFIP Network Operations and Management Symposium (2018).

[23] Buterin, V., Ethereum White Paper: A Next Generation Smart Contract & Decentralized Application Platform (2013).

[24] Meier A., Terán L. (2019) eSettlement. In: eDemocracy & eGovernment. Progress in IS. Springer, Cham.

[25] Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., Holz, T.: We value your privacy... now take some cookies: measuring the GDPR's impact on web privacy. In: Network and Distributed System Security Symposium (NDSS) (2019).

[26] Rhonda Bradley., Data Privacy Concerns: An Overview for 2019, The Manifest (2019).

[27] Mike Isaac, Sheera Frenkel., Facebook Security Breach Exposes Accounts of 50 Million Users, The New York Times (2018).

[28] Dell Cameron., Sacramento Bee Leaks 19.5 Million California Voter Records, Promptly Compromised by Hackers, Gizmodo (2018).

[29] Michael Sykes., Data of 37 million Panera Bread customers may have been exposed, Axios (2018).