Research Directions on Big IoT Data Processing using Distributed Ledger Technology: A Position Paper

Benjamin Agbo, Yongrui Qin and Richard Hill
School of Computing and Engineering, University of Huddersfield, U.K.

Keywords: Big Data, Blockchain, IOTA, Internet of Things.

Abstract: The significant growth and adoption of Internet of Things (IoT) solutions has led to tremendous increase in the generation of data. The need for high speed data processing has become very important to meet with the ever increasing volume and velocity of IoT data, due to the large scale and distributed nature of IoT infrastructure and networks. Present cloud based technologies are struggling to meet up with these needs for real time data processing in the midst of enormous amounts of data. The success of bitcoin has inspired more research in the application of Distributed ledger technologies in various domains. The decentralized nature of these platforms have enabled security and privacy of data in previous research and their architecture has a potential for enabling large scale decentralized data processing. In this paper, we identify some open areas of research in the use of distributed ledger technology and propose a framework for storing, analyzing and ensuring the security of large volumes of IoT data.

1 INTRODUCTION

The Internet of Things can be described as a network of multiple homogeneous and heterogeneous devices that have the ability to sense, process and share data generated from their surroundings (Singh et al., 2018). By enabling easy communication and access with a wide range of devices such as home appliances, sensors, actuators, surveillance cameras, vehicles, etc., IoT will need to deploy more applications that will make use of the potential large amount and variety of data that will be generated by IoT devices (Zanella et al., 2014).IoT cuts across various application domains and several surveys have presented different views of IoT: (Gazis et al., 2015) focused on the challenges of IoT, (Bandyopadhyay and Sen, 2011) identified sets of IoT standards and (Miorandi et al., 2012) presented various applications of IoT. In this paper, we focus on some challenges involved in processing Big IoT data and explore the potential of distributed platforms in enhancing Big IoT data processing. Thousands of use cases and applications of IoT can be identified in each domain, bringing new challenges in the need for interconnection among devices. Various solutions have been proposed to solve the issues associated with the interconnection of smart things. However, the massive generation of data by IoT devices has led to new demands for real time processing of IoT data and improved security and privacy (Gazis et al., 2015). This proposes new challenges, driving research in industry and academia.

The introduction of blockchain technologies has gained significant interest in industry and academia due to the huge success of Bitcoin (Nakamoto, 2008). Blockchain is a distributed ledger system which was developed to enable trust between participants (nodes) in a multi-party business network (Vo et al., 2018). Transactions in a blockchain network can occur without any third party intervention. Blockchain could be viewed as a public ledger where transactions are stored in each block. The size of a blockchain grows as new transactions are continuously added. The key characteristics of this distributed ledger system are: decentralization, anonymity, persistency and auditability (Zheng et al., 2017). Blockchain can be used to provide various solutions and services including smart contracts (Kosba et al., 2016), security services (Noyes, 2016), financial services (Peters et al., 2015). A framework for data sharing using blockchain has been tested in a clinical research conducted by (Benchoufi and Ravaud, 2017). It has also shown huge potential for enhancing security, privacy and trust in a multiparty network. This has inspired various attempts to integrate distributed ledger technologies in the Internet of Things (Ramachandran and Krishnamachari, 2018; Panarello et al., 2018).

Agbo, B., Qin, Y. and Hill, R.
Research Directions on Big IoT Data Processing using Distributed Ledger Technology: A Position Paper.
DOI: 10.5220/0007751203850391
In *Proceedings of the 4th International Conference on Internet of Things, Big Data and Security (IoTBDS 2019)*, pages 385-391
ISBN: 978-989-758-369-8

A framework for processing large volumes of IoT data is proposed in this paper, integrating IOTA technology in the big data process life cycle.

The remainder of the paper is organized as follows. Section 2 provides a view of Big Data. Section 3 lays the identified challenges and motivation for Big IoT data processing. An overview of some distributed ledger technologies is provided in Section 4. Section 5 points out some open research challenges and Section 6 concludes the paper.

2 BIG DATA

The growth in the amount of data generated by IoT devices has played a key role in the big data domain (Jara et al., 2014). Big data can be described as any form of structured, semi-structured or unstructured data characterized by large volume. The wide adoption of IoT has brought new challenges in big data analytics because IoT has brought about requirements for the collection and processing of heterogeneous data obtained from different sensor devices (Marjani et al., 2017). Big data is an important ingredient for making constructive decisions. Most importantly, hidden patterns and deeper understanding of events can be discovered when big data is efficiently mined. Big data is popularly characterized using 4Vs, i.e., Velocity,Volume,Variety and Veracity (Géczy, 2014):

- **Velocity:** Recent technological advancements have resulted in the rapid generation of real-time data from various platforms. This rapid increase in the generation of data is what researchers characterize as the velocity of data.
- **Volume:** Various domains such as business, health, transport, entertainment and other aspects of human life, uses and generates large amount of data in terabytes, exabytes or zettabyte. Big data is characterized by large volume of data.
- **Variety:** The new big data era must make use of fast paced data generated from multiple sources. Data generated from different sources often have heterogeneous formats. The structured, semi-structured and unstructured nature of data describes the variety of big data.
- **Veracity:** This describes the importance of quality data as a tool for drawing accurate conclusions.

It is important to note that, big data in any form (structured, semi-structured or unstructured) requires preprocessing before it can be used for analytics.

2.1 Characteristics of Big IoT Generated Data

According to (Mahdavinejad et al., 2017), data generated by IoT devices can be continuous or dynamic in nature. For instance, data generated from traffic, energy and health management applications would generate large continuous data (Big Data). However, processing data generated in different rates is a challenging task, as they vary from different devices. Another example can be seen in the measurement of data from IoT devices. The frequency of updates from GPS sensors is measured in seconds whereas, temperature sensor updates may be measured hourly. The risk of important information loss always exists regardless of the rate at which data is generated. Another characteristic of IoT data is its dynamic nature. For example, the data generated from sensors embedded in autonomous cars will differ based on locations and time.

It is important that data generated by IoT devices is of high quality. However, different levels of data quality can be noticed in IoT data due to the fact that they are obtained from multiple heterogeneous sources. The quality of information obtained from each data source is dependent on the following factors (Jara et al., 2014):

- Precision of data collection or error in measurement.
- Noise from devices in the environment.
- Discrete measurements and observations.

3 MOTIVATIONS AND CHALLENGES

The high adoption of intelligent devices and the increased use of sensor data, Internet data, finance data, streaming data, etc., has brought about new challenges in the management, analysis and even the usage of big data in the IT industry. In this paper, we consider how we can leverage the potential of distributed ledger technologies in providing more solutions to big data challenges. (Ji et al., 2012) has identified three important aspects that are important in processing big data and we present the challenges around the three points as follows:

3.1 Big Data Management and Storage

Current big data management technologies are struggling to meet up with the needs of big data because the velocity of big data is far greater than the increasing speed of storage capacity. Also, previous computer algorithms find it difficult to directly store data generated from the actual world due to the heterogeneity of real world data. The use of virtual server technologies can also exacerbate the problem when high concurrent I/O is required. This challenges have motivated more investigations into decentralized processing to provide solutions to present master-slave models.

3.2 Big Data Analysis

Speed is very important when processing big data queries. However, this is usually a time consuming process because queries cannot traverse an entire database within a short period of time. Presently, big data indices only process simple data types even though big data requirements are becoming more complicated. Real time pre-processing can help increase query response times. In addition, exploiting the potential of decentralized platforms like IOTA could help target specific queries, thereby improving query response.

3.3 Big Data Security

Various big data applications have helped organizations reduce their IT costs. However, the issue of security and privacy still remains a challenge in big data storage and processing due to the massive involvement of third party services. Unlike traditional security techniques, big data security is concerned with how big data can be processed without exposing sensitive user information. Exploiting the potential of distributed ledger technologies could be a good solution in ensuring security and privacy of big data due to the elimination of third parties on these platforms. Also, the decentralized architecture of these technologies could have significant impact in handling big data challenges.

In the next sections, we will provide an overview of popular distributed ledger technologies and further considerations in data processing.

4 DISTRIBUTED LEDGER TECHNOLOGIES

4.1 Blockchain Technology

Blockchain is an interconnection of decentralized blocks, inspired by the distributed ledger technology. BC enables direct crypto transactions between sellers and buyers without any central authorization. Each peer in a blockchain network records information from each transaction. The records generated from individual pairs form a block (Hassani et al., 2018). The key identifiable characteristics of blockchain are shown below (Zheng et al., 2017)

- **Decentralization:** The integration of technologies such as digital signature, consensus algorithms and cryptographic hash has eliminated the need for a centralized transaction system which required third party intervention in validating transactions.
- **Persistency:** In a blockchain network, transactions can easily be verified and invalid transactions will be rejected by honest miners. It is difficult to rollback transactions once they have been added to a blockchain. This makes it easy to discover transactions that are invalid
- **Anonymity:** Each user in a blockchain network interacts using a generated address which masks the real identity of the user. However, it is important to note that perfect preservation of user information is not guaranteed in blockchain due to the fact that the transactions and balances of each user's public key is visible to the public.
- **Auditability:** Data regarding user balances are stored in a Bitcoin blockchain using the Unspent Transaction Output (UTX-O) model. This model follows a principle whereby each transaction is required to refer to previous unspent transactions. Once new transactions are recorded in a blockchain, the status of previously unspent transactions switches from unspent to spent. This makes it easy to track and verify transactions.

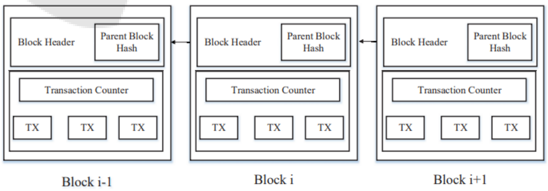

Figure 1: An Example of Blockchain Nodes.

Although research has shown that blockchain technology can add more value to the operations of future IoT systems, this technology is still facing some technical challenges. Firstly, the issue of scalability is a major concern faced with blockchain. This large amount of data generated by blockchain is due to the replication and transmission of data recorded from every transaction that takes place on the blockchain network. Due to this reason, blockchain technology does not scale enough to meet the requirements for real-time processing of fast paced data. Secondly, a research conducted by (Eyal and Sirer, 2018) has shown that miners could receive more revenue than they deserve through selfish mining. In this strategy, selfish miners do not broadcast their mined blocks but the private branch created will only be revealed to the public when certain conditions have been satisfied. The rapid growth of the cryptocurrency market and various research attempt to integrate blockchain in IoT, has led to the development of an alternative technology that outperforms the fundamental blockchain technology. In order to scale to the large requirements of millions of IoT devices, there is a need for a more lightweight and efficient platform that can scale to meet the requirements of IoT devices. This challenge is solved by IOTA's major innovation: The Tangle (Popov, 2016).

4.2 IOTA Technology

IOTA is a novel application of cryptotokens, optimized for Internet of Things micro-transactions. Unlike the heavy and complex bitcoin blockchain, which are designed for various use cases, IOTA is designed to be more lightweight. Hence, the name IOTA was coined out which means something small (Foundation, 2016). One of the visions of IOTA is to enable frictionless payments of small amounts between connected IoT devices. While IOTA was developed to provide solutions to the scalability issues faced by IoT devices, its underlying protocol is potentially applicable in various use cases. Although IOTA is popularly known as a cryptocurrency, recent studies have shown its potential application in various domains such as: the elimination of identity theft using TangleID, location based solutions and the interconnection of cyber-physical systems. The idea of the tangle is based on a concept known to computer scientists as a directed acyclic graph (DAG). DAG is simply an assembly of vertices (squares), interconnected by edges (arrows) (Popov, 2016).

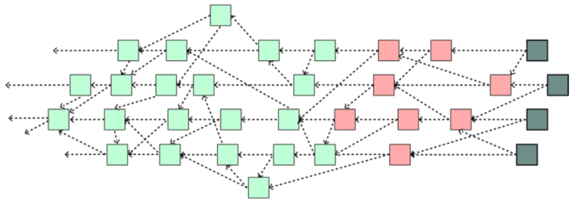

Figure 2: The IOTA Tangle.

The Tangle is a kind of directed graph which forms an IOTA data structure that holds transactions. Every transaction in the graph is represented as a vertex. When a new transaction is added to the tangle, it picks two previous transactions to validate, therefore, introducing two new edges to the directed graph. Unapproved transactions on the other hand are called tips. For example, the gray colored transactions in figure 2 are tips because they have not been validated by any incoming transaction. Every incoming transaction is required to validate a maximum of 2 tips and a minimum of 1tip at random. It is important to note that no rules are imposed on any node with regards to choosing which transaction to validate. However, certain rules can be followed when there is a consensus in the number of nodes following a rule of the same kind (Popov, 2016).

A node must do the following to issue out a transaction (Popov, 2016):

- Choose two transactions to validate according to an algorithm. Ideally, situations may arise where two transactions coincide.
- The node checks if there is any conflict between two nodes and disapproves any conflicting transactions.
- A cryptographic puzzle must be solved before a node can issue a valid transaction. This is similar to the concept used for the bitcoin blockchain.

The storage and transmission of sensor data over an IOTA tangle has been demonstrated using a protocol called masked authenticated messaging. MAM is a layer 2 data communication protocol that allows devices to produce and access encrypted data streams over IOTAs distributed ledger (the tangle) (Handy, 2017). As various transactions are currently recorded using distributed ledger technologies e.g., blockchain, IOTA goes a step further to enhance scalable interaction between IoT devices and thus, it could be used as a platform to enhance some specific data processing concerns.

4.3 An Enhanced Framework for Big IoT Data Processing

As opposed to current commercial Database Management Systems (DBMSs) used for big data processing e.g. Google and Azure, we propose a framework that will ensure better solutions to the bottlenecks caused by peak workloads in processing big data requirements as shown in Figure 3. According to (Ji et al., 2012), scalability and cost are important factors to consider in processing big data. Distributed file systems like Google file system (GFS) and Hadoop Distributed File System (HDFS) were designed and largely optimized for very complex files (measured in gigabytes). The IOTA data market place is currently utilized for a distributed storage of sensor data streams and can be exploited to ensure the verifiability and security of data using MAM. However, as the IOTA platform becomes more data driven, the capability of performing real-time analytics on the IOTA network becomes paramount. A more detailed description of our contribution can be seen in Figure 4.

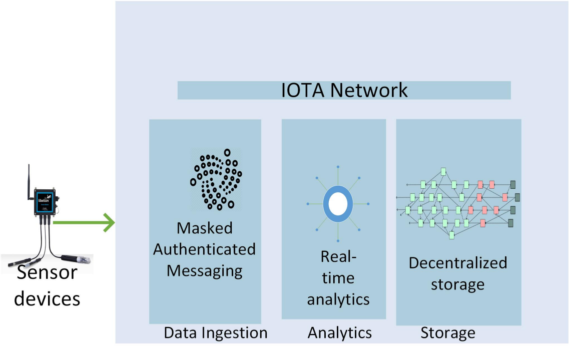

Figure 3: A Framework for Enhanced Big IoT Data Processing using IOTA.

4.3.1 IOTA and Big Data Process Life Cycle

Various phases can be observed in the life cycle of big data processing. The framework in Figure 4, shows the integration of IOTA technology through out the big data life cycle. Various types of IoT sensor devices e.g., temperature detection sensors, proximity sensors, optical alarm sensors, pressure sensors, etc., play a major role in the data generation phase.

The Data generated in the first phase is passed on to the IOTA network through MAM, where an unprecedented amount unstructured, semi-structured and structured data will be collected. This is to ensure that the data collected can only be accessible by nodes that have permission.

Pre-processing of collected data is an integral aspect of the data acquisition phase. Raw data generated by sensor devices often require de-duplication, filtering, compression, integration and cleaning, to handle the issue of inconsistent and incomplete data. Each private sensor node on the IOTA network will have the capability of executing the complete phases of the big data life cycle and connecting with other sensor nodes to aggregate complex data to perform more advanced analytics (Taleb et al., 2015).

The proposed framework will store all processed data as transactions using the IOTA tangle as a database. Each node will have the capability of executing pre-processing algorithms in order to filter the data obtained from sensor devices. The use of IOTA's distributed architecture will eliminate the problem of single points of failure when data is stored and will also help in avoiding scenarios where data processing and storage is controlled by a central node.

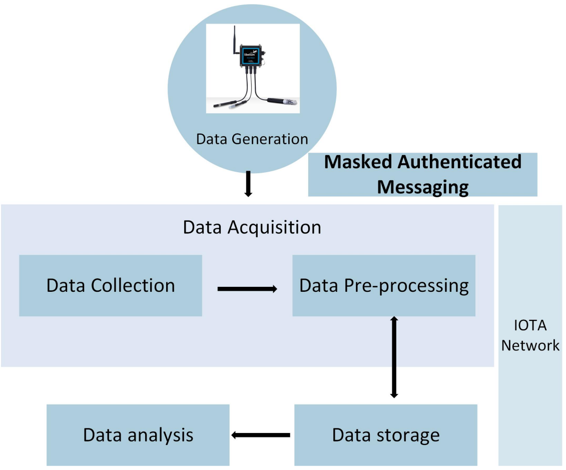

Figure 4: The integration IOTA in the Big Data Process Life Cycle.

As the tangle becomes a popular platform for the secure transmission and storage of future complex IoT data, more requirements will be made to transform these large amount of structured, semi-structured or unstructured data into meaningful information. This can be achieved by exposing sensor data to parallel and distributed processing models such as MapReduce or Spark on the IOTA tangle. When immediate response is required, Spark could be a be a better solution as it can handle fast iterative algorithms and interactive queries. MapReduce on the other hand, spends a significant amount of time fetching data from iterations. In addition, Spark or MapReduce execution nodes can be fused with specific nodes on the IOTA tangle to reduce the need for data transfer before analytics, therefore, enhancing data analytics performance (Vo et al., 2018).

5 OPEN RESEARCH CHALLENGES

Although the concepts and prospects of distributed ledger technologies is evident, its implementation still poses significant amount of challenges. This section identifies some main issues (Reyna et al., 2018).

5.1 Storage Capacity

The ability of distributed platforms to store the amount of data generated in the current big data era has been deeply questioned especially in blockchain. Despite the fact that trust is ensured through the ability of each node to verify transactions, this will lead to the replication of a large amount of data which use up a significant amount of storage.

5.2 Security

One of the well known security challenges facing today's distributed ledger technologies is the denial of service (DoS) attacks, also called sybil or Man in the Middle(MitM). This will obstruct network operations because distributed ledger technologies strongly rely on peer-to-peer communication.

5.3 Legal Issues

The absence of censorship from a central authority is an interesting but yet dangerous peculiarity in distributed ledger platforms. Bitcoin is the first cryptocurrency that rests on a decentralized platform. However, this technology has been accused of promoting fraudulent transactions and illegal conducts. This will require extensive legal considerations in the application of distributed ledger technologies in various domains.

6 CONCLUSION AND FUTURE RESEARCH

The rapid adoption of IoT will continually produce new use cases and requirements. This paper has provided a general view of the nature and characteristics of IoT sensor data. It further identifies resultant big data processing challenges caused by the fast paced generation of data by present IoT devices. The paper further identified the potential of distributed ledger technologies in enhancing big IoT data processing.

The adoption of parallel processing technologies is becoming very important to handle the fast paced generation of data by smart devices today. However, the threat of security and privacy of data still remains and more work is required to eliminate data tampering and ensure the integrity of IoT data.

REFERENCES

Bandyopadhyay, D. and Sen, J. (2011). Internet of things: Applications and challenges in technology and standardization. *Wireless Personal Communications*, 58(1):49–69.

Benchoufi, M. and Ravaud, P. (2017). Blockchain technology for improving clinical research quality. *Trials*, 18(1):335.

Eyal, I. and Sirer, E. G. (2018). Majority is not enough: Bitcoin mining is vulnerable. *Commun. ACM*, 61(7):95–102.

Foundation, I. (2016). Iota development roadmap. (1).

Gazis, V., Goertz, M., Huber, M., Leonardi, A., Mathioudakis, K., Wiesmaier, A., and Zeiger, F. (2015). Short paper: Iot: Challenges, projects, architectures. In *Intelligence in Next Generation Networks (ICIN), 2015 18th International Conference on*, pages 145–147. IEEE.

Géczy, P. (2014). Big data characteristics. *The Macrotheme Review*, 3(6):94–104.

Handy, P. (2017). Introducing masked authenticated messaging. *IOTA Foundation Medium blog*.

Hassani, H., Huang, X., and Silva, E. (2018). Big-crypto: Big data, blockchain and cryptocurrency. *Big Data and Cognitive Computing*, 2(4):34.

Jara, A. J., Genoud, D., and Bocchi, Y. (2014). Big data in smart cities: from poisson to human dynamics. In *Advanced Information Networking and Applications Workshops (WAINA), 2014 28th International Conference on*, pages 785–790. IEEE.

Ji, C., Li, Y., Qiu, W., Jin, Y., Xu, Y., Awada, U., Li, K., and Qu, W. (2012). Big data processing: Big challenges and opportunities. *Journal of Interconnection Networks*, 13(03n04):1250009.

Kosba, A., Miller, A., Shi, E., Wen, Z., and Papamanthou, C. (2016). Hawk: The blockchain model of cryptography and privacy-preserving smart contracts. In *2016 IEEE symposium on security and privacy (SP)*, pages 839–858. IEEE.

Mahdavinejad, M. S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., and Sheth, A. P. (2017). Machine learning for internet of things data analysis: A survey. *Digital Communications and Networks*.

Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A. T., Siddiqa, A., and Yaqoob, I. (2017). Big iot data analytics: architecture, opportunities, and open research challenges. *IEEE Access*, 5:5247–5261.

Miorandi, D., Sicari, S., De Pellegrini, F., and Chlamtac, I. (2012). Internet of things: Vision, applications and research challenges. *Ad hoc networks*, 10(7):1497–1516.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

Noyes, C. (2016). Bitav: Fast anti-malware by distributed blockchain consensus and feedforward scanning. *arXiv preprint arXiv:1601.01405*.

Panarello, A., Tapas, N., Merlino, G., Longo, F., and Puliafito, A. (2018). Blockchain and iot integration: A systematic survey. *Sensors*, 18(8):2575.

Peters, G., Panayi, E., and Chapelle, A. (2015). Trends in cryptocurrencies and blockchain technologies: a monetary theory and regulation perspective.

Popov, S. (2016). The tangle. *IOTA Whitepaper*, pages 1–28.

Ramachandran, G. S. and Krishnamachari, B. (2018). Blockchain for the iot: Opportunities and challenges. *arXiv preprint arXiv:1805.02818*.

Reyna, A., Martín, C., Chen, J., Soler, E., and Díaz, M. (2018). On blockchain and its integration with iot. challenges and opportunities. *Future Generation Computer Systems*.

Singh, M., Singh, A., and Kim, S. (2018). Blockchain: A game changer for securing iot data. In *Internet of Things (WF-IoT), 2018 IEEE 4th World Forum on*, pages 51–55. IEEE.

Taleb, I., Dssouli, R., and Serhani, M. A. (2015). Big data pre-processing: A quality framework. In *2015 IEEE International Congress on Big Data*, pages 191–198. IEEE.

Vo, H. T., Kundu, A., and Mohania, M. K. (2018). Research directions in blockchain data management and analytics. In *EDBT*, pages 445–448.

Zanella, A., Bui, N., Castellani, A., Vangelista, L., and Zorzi, M. (2014). Internet of things for smart cities. ieee internet of things journal, 1 (1), 22–32.

Zheng, Z., Xie, S., Dai, H., Chen, X., and Wang, H. (2017). An overview of blockchain technology: Architecture, consensus, and future trends. In *Big Data (BigData Congress), 2017 IEEE International Congress on*, pages 557–564. IEEE.