Blockchain Data Management and Analytics

Sumit Patel,
Student,
Computer Engineering dept,
GCET, VallabhVidyanagar,

Nihaal Subhash,
Student,
Computer Engineering dept,
GCET, VallabhVidyanagar,

Tirth Patel,
Student,
Computer Engineering dept,
GEC, Gandhinagar,

Abstract

Blockchain technology has emerged as a primary enabler for verification-driven transactions between parties that do not have complete trust among themselves. Bitcoin uses this technology to provide a provenance-driven verifiable ledger that is based on consensus. Nevertheless, the use of blockchain as a transaction service in non-cryptocurrency applications, for example, business networks, is at a very nascent stage. While the blockchain supports transactional provenance, the data management community and other scientific and industrial communities are assessing how blockchain can be used to enable certain key capabilities for business applications.

We have reviewed a number of proof of concepts and early adoptions of blockchain solutions that we have been involved spanning diverse use cases to draw common data life cycle, persistence as well as analytics patterns used in real-world applications with the ultimate aim to identify new frontier of exciting re-search in blockchain data management and analytics. In this paper, we discuss several open topics that researchers could in-crease focus on: (1) leverage existing capabilities of mature data and information systems, (2) enhance data security and privacy assurances, (3) enable analytics services on blockchain as well as across off-chain data, and (4) make blockchain-based systems active-oriented and intelligent.

1. INTRODUCTION

Blockchain (a.k.a. distributed ledger) is an emerging plat- form that is designed to support transactions services within a multi-party business network, with the goal of enabling significant cost and risk reductions for all parties through the creation of innovative new business models. Data maintained within the distributed ledger can only be accessed through the execution of a smart contract (i.e., a stored procedure call on the distributed ledger) that describes rules that govern a transaction. In addition, the design of blockchain technology ensures that no one business entity can modify, delete, or even append any record to the ledger without the consensus from other business entities in the network, making the system useful for ensuring the immutability of data and legal documents.

Given the aforementioned important features, blockchain technology has taken the world by storm in the recent years for its promise to transform every industry. For instance, it has started to be used in a wider range of applications, e.g., Inter-net of Things (IoT). A blockchain enables IoT devices to send data for inclusion in a shared transaction repository with tamper-resistant records, and enables business parties to access and supply IoT data without the need for central control and management. Blockchain for IoT can optimize supply chains by tracking objects as they traverse the export/import supply chain while enforcing shipping and expediting incremental payments.

Similarly, blockchain technology also has the potential to disrupt insurance industry. It has been used in a car micro-insurance application to enable the concept of pay-as-you-go insurance. This application allows drivers who rarely use cars to only pay insurance premium for particular trips rather than the hefty yearly premium. By transparently storing on blockchain all the data pertaining to the actual trip and premium payment, every party in the insurance contract including the driver, the insurance company, and the financial institution is confident that the data are tamper-proof and traceable. This guarantees that any insurance claim request regarding to a trip can be processed quickly and indisputably, hence offering a better customer experience. Further, as the micro-insurance application also requires accessing multiple risk analytic databases such as past driving behavior statistics and past vehicle runtime statistics for computing premiums, a system architecture that allows for maintaining and analyzing both on-chain and off-chain data was also proposed.

In fact, Gartner's 2016 research report1 identified blockchain as one of the key platform-enabling technologies to track. Nonetheless, while there is currently no standard in the blockchain space, there is a growing consensus that blockchain is entering its peak of inflated expectations. The report anticipated that it would take 5 to 10 years for blockchain technology to get mainstream adoption. Further, most of nowadays blockchain efforts, especially when applied to business environments, are still in a nascent state. Research perspectives and challenges related to blockchain have been presented in, but they are mainly for cryptocurrencies and public blockchain environments. The time is ripe for database community to get more deeply involved in solving open problems pertaining to data management and analytics in a permissioned blockchain network for business applications. As the building blocks of blockchain include some combination of database, transaction, encryption, consensus and other distributed system technologies, it is natural to investigate if it is possible to utilize existing capabilities of mature data and information systems through robust integration into blockchain systems. There exist open research issues such as multi-storage and index support, novel transaction concurrency model, scalable transaction throughput, master and reference data management, smart contract management, data security and privacy assurances, as well as information leakage prevention.

Furthermore, even though the blockchain database is useful for transparent persistence of streaming business data, there is no one-size-fits-all database solution for an application. While blockchain is originally designed to maintain transaction data, there is a growing interest in providing analytics capabilities in blockchain-based data systems. Particularly, in this paper we shall elaborate on specific research problems such as built-in analytics for blockchain, and data integration and analytics across on-chain and off-chain data.

Last but not least, with recent advances in areas such as information retrieval, machine learning, and AI, there is a tremendous opportunity to bring cognitive capabilities, e.g., understanding, learning, and contextual awareness into blockchain-based data systems so as to make them active-oriented and intelligent. We shall discuss open research issues encountered while developing intelligent blockchain-based data systems. Our list is by no means comprehensive and other research opportunities exist as well.

2. BACKGROUND

Blockchain:

Blockchain technology provides a framework for building a distributed ledger that can provide consensus, provenance, immutability and finality of transaction data. The use of blockchain was first popularised by Bitcoin, which is a cryptocurrency. In a blockchain, a group of ledger entries, i.e., a list of transactions, are periodically accumulated into a block which contains a cryptographic hash of the prior block linking the blocks together. This way of chaining the blocks allows the global order of the ledger entries to be established and to verify that the content of a particular block have not been modified. Every node in this distributed system maintains its own copy of the blockchain and participates in an appropriate consensus mechanism to keep the replicated data in sync across nodes. For example, Bitcoin uses a consensus protocol called "Proof-of-Work", whereas Hyperledger Fabric develops a variant of Byzantine fault-tolerant (BFT) state machines.

Permission/private blockchain:

In Bitcoin, a public implementation of blockchain, entities that participate in the transfer of assets are anonymous and any entity can participate. In contrast, many business networks may have a need for a distributed ledger that is only accessible to a closed community of known entities. Permissioned blockchain technologies such as Hyperledger Fabric and R3 Corda have been developed to support these requirements, i.e., entities participating in the net-work are identified so that their permissions can be determined and the activities of an entity are only visible to those participants of the business network that have a need to know.

Smart contract:

Some distributed ledger technologies sup-port an additional capability called a smart contract, which is similar to the concept of stored procedure in classical relational databases to some extent. Smart contracts allow the shared business processes within a business network to be standardised, automated and enforced via computer programs to increase the integrity of the ledger.

3. BLOCKCHAIN DATA MANAGEMENT

3.1 Leverage capabilities of mature data and information systems

Multi-storage and index support:

Most blockchain platforms such as Ethereum adopt key-value data model, while a few of them like R3 Corda use relational data model. This makes any single blockchain platform not suited for different types of data used in a wide range of business applications. For ex-ample, geo-location data recorded from vehicles in a car micro-insurance application as discussed in Section 1 may not be efficiently queried using a key-value store. Furthermore, even though blockchain platforms such as Hyperledger Fabric opt for pluggable storage model, developer users have to decide at development time which storage to use, e.g., either Level DB (key-value store) or CouchDB (document store). Therefore, novel techniques are needed for supporting multiple types of data stores such as key-value, document, SQL and spatial data stores simultaneously in the same blockchain system.

Additionally, blockchain is originally not designed to store digital documents, which, however, are a popular type of data shared in a business network as observed in a majority of blockchain solutions that we have been involved. These digital records are usually large, and their aggregate size grows significantly over time. It is infeasible to store these data directly on the blockchain due to several constraints such as storage size, bandwidth and trans-action throughput. One possible solution is to store these records in a third-party offline storage, and maintain their locations and a digital hash of the data on the blockchain for verification. Nevertheless, this approach requires integration of blockchain and offline storages. Thus, it is critical to develop blockchain systems with built-in offline storage strategies for handling big data.

We also observe from the implementation of those blockchain solutions that rich queries (e.g., conditionals, operators etc.) of data on blockchain are typically read-only and based on non-primary keys. To deal with these situations, explicit smart con-tracts for maintaining secondary indices have to be developed. This motivates exciting research problems related to index management in blockchain-based data systems.

Master data management:

Unlike blockchains used in public cryptocurrency environments, a business blockchain network is not a single universal collaborative environment for every organization to join in this same network. Instead, each network usually includes a specific set of organizations sharing some common business interests, and more importantly, an organization may join a number of different blockchain networks due to the large scope of their business. It is likely that each network will have a different data schema and may record a different version of some common data referring to the same entity across the networks. Therefore, organizations need master data management rules, processes and techniques in order to consolidate data across multiple blockchain networks that they participate in. In addition, we also envision an interesting opportunity for future research to explore a new concept of cross-chain smart contracts that run across multiple blockchains.

Reference data management:

Since the data in blockchain cross over the boundary of organizations, semantically right interpretation of data is must. Hence, one important problem is interpreting the data w.r.t. reference data and business glossary. Particular technical problems include identification of reference data entities, automatic interpretation/conversion, and managing the reference data as they are provided by external sources. Another technical question is whether this logic of references should be handled in the smart contract or at application level. Further, query processing on blockchain must take such context, i.e., references, into account to carry out meaningful processing.

Scalable transaction throughput:

There continues to bethe quest for scalable transaction throughput in blockchain. The blockchain in Bitcoin uses “Proof of Work” (PoW) consensus method that is computationally expensive (by design) for having to solve a cryptographic puzzle in the process. Instead, the permissioned blockchains where participants are identified use consensus methods based on variants of Byzantine fault-tolerant (BFT) state machines, which have been chosen to provide higher transaction throughput and lower consensus latency as shown in the recent benchmark. Nevertheless, even with that performance improvement, the benchmark paper concludes that current blockchain technologies are not suited for large-scale data processing workloads. Consequently, there still exists theneed of novel methods, e.g., implicit consensus and sharding data that provide high levels of transaction throughput for blockchain. A different direction to achieve scalable throughput, e.g., BigchainDB, is developing blockchain-like trusted trans-actions on top of existing modern distributed database systems.

ACID properties:

Presently, no blockchain platform fully guarantees ACID properties because blockchain is not de-signed to support databases, nor it is always tractable to support these properties on distributed ledgers. Nevertheless, as we apply transactional semantics to blockchain and use it for data management, we need to assure the important ACID properties. It is interesting to see that even though blockchain platforms are designed to maintain transaction data, they adopt a simple concurrency control, or not at all, to deal with concurrent trans-actions accessing the same data item. Transactions in blockchains are validated based on a first-come first-served basis. That is, the first transaction to get endorsed and committed by all nodes in the blockchain network wins and invalidates other conflicting transactions that are concurrently modifying some common data items. In this case, all other client applications executing those conflicting transactions have wasted time waiting for the notification from the blockchain about the rejected status of their submitted transactions. Hence, an important research issue here is to develop novel concurrency control models for blockchains so that they are applicable to a wider range of applications.

3.2 Enhance information protection

Blockchain-enabled applications involving security- and privacy-sensitive data, e.g., financial and healthcare, requires confidentiality, security and privacy assurances at different levels to be supported by the system, which are mandated by regulatory compliance requirements such as HIPAA and GDPR.

Confidentiality:

Access control mechanism is mainly used to protect access data on permissioned blockchains. However, access control is insufficient to provide protection from data exposure. Data that is stored on distributed ledger of the blockchain networks need to be encrypted. Thus, it is essential to determine the “computational hop” in which data shall remain encrypted, primarily because if we assume that data shall remain encrypted across its life cycle, processing of such data using smart contracts shall be difficult (unless we use fully homomorphic encryption or some form of malleable encryption schemes). Querying data on blockchain has to be enabled even when the data is encrypted. Models such as “search on encrypted databases” if implemented shall have severe impact on the performance of the blockchain system. Therefore, in the presence of encrypted data available for querying, the query execution system on blockchain has to be properly designed and implemented.

Privacy:

Bitcoin and Ethereum claim to support some form of privacy for transactional information, which, however, has been shown not to be entirely privacy-preserving. Data on blockchain may need to be shared across others for analytics. There are use cases where research needs to be carried out on implementing “right-to-forget” on blockchain. GDPR and EU regulation on data privacy have recently asked Google and other internet companies to support “right-to-forget” – a user may ask the service providers to remove their data from the search results or from the system altogether. When a blockchain stores healthcare, finance and such other sensitive data, a user may request the blockchain provider to delete some or all the records pertaining to the user, which is hard to support today due to the immutability of blockchain. This necessitates cryptographic techniques to rewrite history in blockchains.

Information leakage prevention:
For blockchains that maintain unstructured data such as business documents, a more comprehensive data redaction mechanism would be needed to protect business-critical information contained in these documents from unintentional disclosure. Typically, these documents can be accessed fully or they are protected completely. However, there is real need that documents can be released partially by redacting certain data entities that should be prohibited from the users. Thus, it is important to support fine-grained access control on these documents (i.e., access control at data entity level rather than at document level). This fine-grained access control can be achieved by detecting and removing the business-critical information from the document shared on the blockchain based on the role of the user who requests to access the document.

3.3 Manage smart contracts

Smart contract governance:
As smart contracts capture the shared business processes between parties in a blockchain net-work, the governance of these smart contracts at every step of their life cycle including business analysis, design, development, testing, deployment, and monitoring would in general require the coordination and approval of these multiple parties. Nevertheless, this governance process is not standardized and automated as per current practice. Instead, it is common that only one or a subset of parties are delegated and trusted to manage the entire life cycle of a smart contract. This points to the need of tools that allow automated and collaborative governance of smart contracts.

Smart contract template:
As current practice, smart con-tracts are manually programmed by developers after studying requirements described in legal documents agreed by multiple par-ties. However, this process is time consuming and error-prone as there is currently no standard regarding the legal enforceability of code-based contracts. It is, therefore, important to make the development of smart contracts as much automated as possible. An open research issue in this direction is to define a standardised semantic framework for smart contracts that considers both operational and non-operational aspects based on existing legal documents. Another related research challenge is to propose standardised templates for automated generation of legally-enforceable smart contracts from legal documents. This requires understanding natural language used in legal documents and identifying operational parameters that can be used as the connection between legal agreements and smart contracts.

Trusted smart contracts:
Smart contracts vulnerable to at-tacks can expose blockchain data or contain backdoors that may be used to exfiltrate data from the blockchain. Hence, it is essential to develop security analysis technology for smart contracts and reasoning about their semantic trustworthiness, i.e., trust between parties defined based on the semantics of smart con-tracts rather than the digital signatures of the code for smart contracts. Several open questions need to be addressed, e.g., how to detect bugs in smart contract codes, how to deal with smart contract codes behaving erratically and how to update with correct codes without impacting the network.

4. BLOCKCHAIN DATA ANALYTICS

4.1 Built-in analytics for blockchain:

As the original blockchain is purely a transaction repository, an execution engine will be required for analytics running directly on blockchain data. A possible solution to this problem is to make blockchain data readily accessible by data parallel processing systems such as MapReduce or Spark. In particular, an input reader could be implemented so that MapReduce and Spark programs are able to scan through blockchain data efficiently. Further, MapReduce or Spark execution nodes can be physically co-located with blockchain data nodes to reduce the need of data transfer, and hence improving analytics performance. Apart from the above batch analytics, there are also use cases such as IoT applications in the supply chain domain as discussed in Section 1 in which lightweight or edge analytics capability (i.e., analyzing data as ingested into blockchain) would be critical to the system.

4.2 Integration and analytics across on-chain and off-chain data:

It is worth noting that the need of data integration across multiple blockchains that an organization participates in, as discussed in Section 3.1, is just one dimension of the problem. Another dimension of data integration problem comes from common data entities referred by both the blockchains and the organization’s legacy systems of record. In particular, whereas blockchains function independently of legacy systems in most cases, at some point in the application development process organizations will need to integrate blockchain data with their existing systems of record for deriving complete business insights. Since multiple parties are joining a blockchain network, cases of overlapping or inconsistent data between the blockchain and their legacy systems will likely arise. As a consequence, there is much scope for development of new techniques in entity resolution for big data spanning across blockchain and off-chain data.

In addition, as the analytics now spans across on-chain and off-chain data systems, query processing over federated data and optimisation techniques would be the key to the performance of query federation. For example, would the strategy that exports all data on blockchain into an off-chain database where all the analytics are executed be optimal? In contrast, are there better approaches that only materialize part of relevant blockchain data in the off-chain database and how it could be done dynamically given the changing workload? More importantly, it is also challenging to ensure the immutability of the data that has been exported from blockchain into external data stores. These are very interesting research issues and they require more exploration on query federation, translation, and optimization, as well as data security in the context of analytics over both on-chain and off-chain data.

5. INTELLIGENT BLOCKCHAIN SYSTEMS

As discussed in Section 1, blockchain technology has started to be used in a wider range of applications, e.g., Internet of Things (IoT). Nevertheless, the volume of data generated in this era of the Internet of Things is growing significantly, which puts blockchain systems to their limits of transaction throughput and storage capacity. Consequently, when a new piece of data arrives, it is important for the blockchain system to be able to understand the input data, reason about its relevance to the business so as to determine whether dropping the data or accepting and storing it in the blockchain. Recently, a technique to reduce data acquisition cost by only accepting data that is useful for answering queries has been proposed. However, none of prior systems is able to self learn the relevance of incoming data to the business. This necessitates an active-oriented and intelligent blockchain system for making sense and intelligently classifying incoming data, which greatly helps reduce redundant data storage and computation at later stages. In fact, intelligence can be embedded at every step in the pipeline of data processing inside a blockchain-based system, similar to the concept of intellective data warehousing.

6. CONCLUSIONS

We have highlighted research topics that characterize the com-mon issues of on-going data management and analytics problems encountered in the development of real-world blockchain applications. We hope that this study could provide a basis for further research to identify likely solutions to these open problems.

7. REFERENCES

[1]. 2008. Bitcoin: A Peer-to-Peer Electronic Cash System. https://bitcoin.org/ bitcoin.pdf. (2008)

[2]. 2013. Health Insurance Portability and Accountability Act of 1996 (HIPAA). https://www.hhs.gov/hipaa/for-professionals/index.html. (2013).

[3]. 2016. Directive 95/46/EC (General Data Protection Regulation). http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679. (2016).

[4]. 2017. BigchainDB. https://www.bigchaindb.com. (2017).

[5]. 2017. Corda. https://github.com/corda/corda. (2017).

[6]. 2017. CouchDB. http://couchdb.apache.org. (2017).

[7]. 2017. Ethereum. https://www.ethereum.org. (2017).

[8]. 2017. Hyperledger. https://www.hyperledger.org. (2017).

[9]. 2017. LevelDB. https://github.com/google/leveldb. (2017).

[10]. Giuseppe Ateniese, Bernardo Magri, Daniele Venturi, and Ewerton Andrade. 2017. Redactable Blockchain - or - Rewriting History in Bitcoin and Friends. In *Proc. of IEEE European Symposium on Security and Privacy*. 111–126.

[11]. Marcella Atzori. 2017. Blockchain-Based Architectures for the Internet of Things: A Survey. https://ssrn.com/abstract=2846810. (2017).

[12]. Joseph Bonneau et al. 2015. SoK: Research Perspectives and Challenges for Bitcoin and Cryptocurrencies. In *Proc. of IEEE SSP*. 104–121

[13]. Christopher Clack et al. 2016. Smart Contract Templates: foundations, design landscape and research directions. *CoRR* abs/1608.00771 (2016).

[14]. Jeffrey Dean and Sanjay Ghemawat. 2004. MapReduce: Simplified Data Pro-cessing on Large Clusters. In *Proc. of OSDI*. 10–10.

[15]. Prasad Deshpande et al. 2015. The Mask of ZoRRo: preventing information leakage from documents. *KAIS* 45, 3 (2015), 705–730.

[16]. Tien Tuan Anh Dinh et al. 2017. BLOCKBENCH: A Framework for Analyzing Private Blockchains. In *Proc. of SIGMOD*. 1085–1100.

[17]. A. Kosba et al. 2016. Hawk: The Blockchain Model of Cryptography and Privacy-Preserving Smart Contracts. In *Proc. of IEEE SSP*. 839–858.

[18]. Danny Yang et al. 2017. Survey of Confidentiality and Privacy Preserving Technologies. (2017). R3 Research.

[19]. Eleftherios Kokoris-Kogias et al. 2017. OmniLedger: A Secure, Scale-Out, Decentralized Ledger. Cryptology ePrint Archive, Report 2017/406. (2017).

[20]. Hoang Tam Vo et al. 2017. Blockchain-based Data Management and Analytics for Micro-insurance Applications. In *Proc. of CIKM*. 2539–2542.

[21]. Mukesh Mohania et al. 2018. Active, Real-Time, and Intellective Data Warehousing. In *Encyclopedia of Database Systems*. 1–10. To appear at https://doi.org/10.1007/978-1-4899-7993-3_8-3.

[22]. Zhijie Ren et al. 2017. Implicit Consensus: Blockchain with Unbounded Throughput. *CoRR* abs/1705.11046 (2017).

[23]. Michael Fleder et al. 2014. Bitcoin Transaction Graph Analysis. (2014).

[24]. Theo Haerder and Andreas Reuter. 1983. Principles of Transaction-oriented Database Recovery. *ACM Comput. Surv.* 15, 4 (1983), 287–317.

[25]. Zheng Li and Tingjian Ge. 2016. Stochastic Data Acquisition for Answering Queries as Time Goes by. *PVLDB* 10, 3 (2016), 277–288

[26]. Loi Luu et al. 2016. Making Smart Contracts Smarter. In *Proc. of ACM CSC*. 254–269.

[27]. Bill Marino and Ari Juels. 2016. *Setting Standards for Altering and UndoingSmart Contracts*. 151–166.

[28]. C. Mohan. 2017. Blockchains and Databases. *PVLDB* 10, 12 (2017), 2000–2001.

[29]. Pablo Seijas et al. 2016. Scripting smart contracts for distributed ledger tech-nology. *IACR Cryptology ePrint Archive* 2016 (2016), 1156.

[30]. Michael Stonebraker and Ugur Cetintemel. 2005. One Size Fits All: An Idea Whose Time Has Come and Gone. In *Proc. of ICDE*. 2–11.

[31]. Florian Tschorsch and Bjorn Scheuermann. 2016. Bitcoin and Beyond: A Technical Survey on Decentralized Digital Currencies. *IEEE CommunicationsSurveys and Tutorials* 18, 3 (2016), 2084–2123.

[32]. Marko Vukolić. 2016. *The Quest for Scalable Blockchain Fabric: Proof-of-Workvs. BFT Replication*. 112–125.

[33]. Xiao Yue et al. 2016. Healthcare Data Gateways: Found Healthcare Intelligence on Blockchain with Novel Privacy Risk Control. *Journal of Medical Systems* 40, 10 (2016), 1–8.

[34]. Matei Zaharia et al. 2012. Resilient Distributed Datasets: A Fault-tolerant Abstraction for In-memory Cluster Computing. In *Proc. of NSDI*. 2–2.