A quick overview about challenges in Data Sciences and Blockchain

Thomas Suau

July 16, 2023

Abstract

The goal of this paper is to give a brief overview about how data knowledges can be used in Blockchain research.

The goal is not to solve actuals problems only to give some research proposals in Blockchain and Data Sciences.

With recent developments over Bitcoin protocol [28] and ethereum scaling proposals with layer 2 [31] data is massively arriving in the Blockchain. This will issue numbered of new problems and new challenges.

I hope to clarify the answer about : how data and blockchain are related ?
I want to focus on what interesting problems can be delve into the topic of blockchain and data. As generally as possible.
To stay synthetic I kept 5 major challenges and I purpose a simplified plan to attack them.

Contents

1 Introduction

As we know data are today a core in major part of computer sciences. Without doing again the history of computer science the data is one of the key concept in information theory. With the growth of computer capacities we are today able to store and manage huge amount of data. This is so called big data [10]. These data are from different forms and often stored in dedicated server. One very common language to request these data is `SQL` even though many new requesting technics appearing called `No-SQL` with Mongo DB[7], Cassandra DB[12], GraphQL [9] and many others [25].

In another hand, since 2008 there is a new technology deployed : the Bitcoin [21], with in its core the blockchain. The blockchain technology allow us to share value across digital realm. With Ethereum in 2014 [2], a new type of blockchain appeared : a decentralised application platform. The idea behind Ethereum is to get everyone access to a decentralised Virtual Machine (the Ethereum Virtual Machine [11]). We can use the EVM if we are paying gas [1] in ETH (the cryptocurrency associated to Ethereum Blockchain). The access to EVM means that we have access to an hardware and we can make action on this hardware.

In this point of view, the blockchain technology allow us to share our hardware and to be paid for this. We are also able to execute code (more precisely bytecode [22]) and store data into the EVM storage (please find the schema in Appendix 3).

What are issues and challenges associated to those data ?

This article assumes to give a brief overview about actual problem in data and blockchains. It wouldn't be exhaustive because of the complexity of the topic should drive us around deep research as shown in Deepa, Pham et alt. [6] :
"Despite many research efforts, we are not aware of any survey that comprehensively studies the applicability of blockchain for big data applications. Although the survey in [32] reviews blockchain for big data applications and challenges, it is very short and not updated since it has been published several years ago."

Data are still in a human unreadable format, difficult to query as assume in Prytarski, et alt. [23] and European Blockchain Service Infrastructure (EBSI) is far to be usable.

Many developments should be done in this way and as it's assume in Deepa, Pham et alt. [6] : "Blockchain with its decentralization and security nature has the great potential to improve big data services and applications".

2 State of the art

The state of the art will be brief as the research in this field. There are not many qualitatives research around blockchain compare to challenges blockchain is facing.

There is Hassani, et alt. [10] published in 2018. This article important for this field, is not at the cryptography academic level. There is a lack of technical aspect

and no authors are doing computer sciences in their research field. It's mainly general purpose and blockchain is apparently misunderstood in this paper. We can confirm this by checking their link to explain what blockchain is from a confused picture in 2.1. There is a book published in 2015 Handbook of digital currency [**?**]. This book is an aggregation of different authors and topics around blockchain and data. The main focus is rather finance than blockchain or data themselves. It's very large and not as specific as current challenges.

There are many references about specific applications like data integrity in Lui, et alt. [19], securing trading data in Wai, et alt. [5] , fraud detection in Kamisalic et alt. [14] and others like in medical records in Liu, Lam et alt. [20] or supply chain in Wang, et alt. [27].

The only academic paper with approaches, challenges and future perspective is about Blockchain for Big Data in Deepa, Pham et alt. [6]. Unfortunately the article conclude with "Despite many research efforts, we are not aware of any survey that comprehensively studies the applicability of blockchain for big data applications. Although the survey in [32] reviews blockchain for big data applications and challenges, it is very short and not updated since it has been published several years ago.".

A deeper analysis of Prytarski, et alt. [23] should be done. Even though their experiment of SQL queries is very short, they provided a good bibliography in their state of the art that should be interesting to delve.

With the recent update of Ethereum client `geth` [8], we are now able to make GraphQL[9] queries. This is also a major new updates and we need to update scientific research according to this.

What I'm purposing is to give another regard to the link between Data and Blockchain.

3 Challenges

1. **Retrieve data stored on-chain**
   There exist tools to make `SQL` request Access Bitcoin and Ethereum open datasets for cross-chain analytics provided by AWS. The query processing on blockchain is still a challenge as assume in Przytarski, Stach, Gritti et alt. [23] where they can only try some `SELECT` statement by migrating original Ethereum LevelDB in SQLite. Regarding the exponential growth of the Ethereum blockchain, we need to think about other ways to do so. The creativity will be a big challenge too.
   Recent updates in `geth` [8] client haven't been regarded so far by academic community and it can be a good starting point.
   As an example of data in a blockchain you can find one Bitcoin transaction which contains the JSON file below (more details about this is given Appendix 5).
2. **Data visualisation for on-chain data**
   This is a real problem in blockchain ecosystem. The Blockchain Data Platform

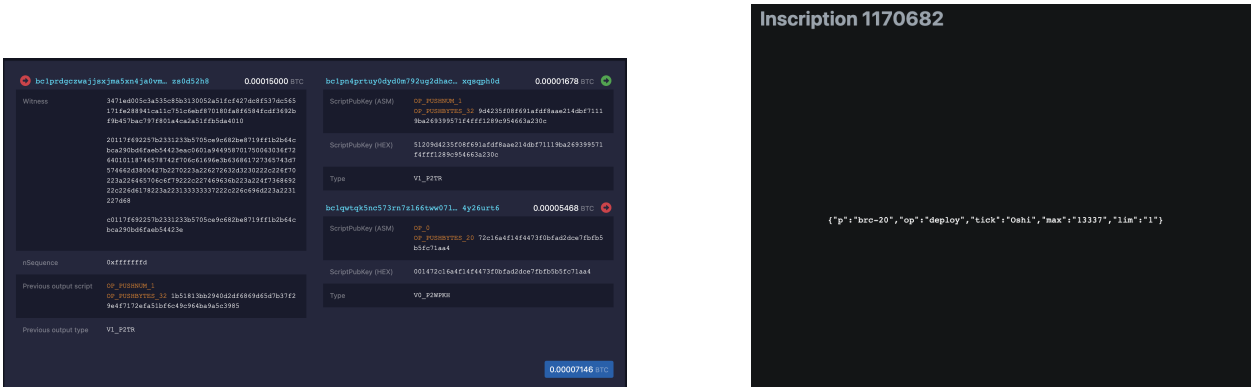

Figure 1: Bitcoin transaction (left) data inscribed (right)

- Chainalysis makes several use of on-chain data visualisation. One great article that could be reviewed in this way is [26] which aims to give every tools to provide such visualisation. The master thesis of M. Kaandorp [13] should also be regarded for its querying technics.
A basic example of what we can do is `mempool.space`. In the Decentralised Finance (DeFi) space data aggregation as DeFi Lama is very important. Professional tools can be used to create such aggregation about different protocols and with different utilities. Make a visual representation of these data can be a game changer in the blockchain realm.
With the development of European Blockchain Service Infrastructure (EBSI), it can be a major problem for future industries and companies.

3. **Market data**
This is obviously a big deal for financial market especially in the case where this data are more or less easily available. There is a recent article which considered wash trade detection from on-chain liquidity data [4]. This would begin to be a big research field when we know that Blackrock is arriving into the Bitcoin Market Nasdaq refiles BlackRock's bitcoin ETF application with SEC.
One major advantage of these data are to be more convenient than rough blockchain data [30]. As shown in Lie, Xie et alt. [29] we can make implementation of deep learning to use more efficiently market data from blockchain.

4. **Data from mining**
Mining data have different aspects.
In one hand it's to mesure Bitcoin electricity consumption [18, 16] Cambridge Bitcoin Electricity Consumption Index (CBECI).
In another hand, it can be with many different use case for example in Zhu, Liu et alt. [33] they present the interest "to de-anonymize" transactions with making use of IP address from Bitcoin miners. The security of the network is depending on HashRate power [3] and we can link this data with market data as in Kubal et alt. [15] to learn more about the network.

5. **Store and authenticate data**
   The problem of data integrity is a real challenge in big data [6]. I need to learn more about data integrity to be able to give arguments about how the blockchain could help in this way. But in [6] we can find an array which summarises challenges that blockchain could solved in Big Data management.

Ref	Cloud based services	Challenges faced by Big data	Solutions provided by Blockchain
[65]	Data collection.	Data collection is exposed to various malicious attacks and threats.	Blockchain provides energy efficient data collection and secure data sharing environment using Ethereum.
[66]	Data transmission/sharing.	Lack of authorization for data sharing in edge nodes and response time is more.	Blockchain based futile transaction filter algorithm helps to access data from cache layer instead of storage layer and helps to reduce response time and storage overhead. Smart contracts are used for authorization.
[68]	File storage system.	Unauthorized access to the electronic file system. Privacy, security and redundancy problems.	Blockchain integrated with IPFS provides the solution by implementing decentralized platforms to solve file redundancy problems and provides security to the file storage system. Hash value of data is stored in blockchain to provide authenticity to the users and an attribute based encryption method is applied before data storage in cloud.
[69]	Database management system.	Data stored in distributed database is exposed to internal and external attacks.	Blockchain overcomes data tampering using time stamping method. Virtual shared ledger is applied to store the transaction history. Database transactions are recorded in block and each block is interconnected with each other using cryptographic hash value. Blockchain based solution integrates storage servers, cryptographic algorithms for a reliable database access.
[71]	Data training/learning process.	Various entities share data to integrate the dataset with various attributes and train the ML classifier. Data privacy issue occurs while sharing data from various entities.	Blockchain consortium and homomorphic cryptosystem provide a secure training platform without the intervention of a trusted third party. Blockchain provides a secure environment for communication between the entities.
[76]	Data privacy preservation.	User privacy is an issue in digital scenarios in big data era. Services provided by third parties are exposed to security breaches and data misuse.	Blockchain provides immutable, verifiable and decentralized ledger to record the transactions in digital scenarios.It provides facilities to the user to control their personal data. Crypto-privacy methods are applied to solve privacy preserving problems.

Figure 2: SERVICES PROVIDED BY CLOUD ENVIRONMENT FOR BIG DATA, CHALLENGES AND BLOCKCHAIN BASED SOLUTIONS.

This list is not exhaustive and show what paths can be taken in data blockchain related problems.

4 Some paths to do this ?

It can be done in different ways but I think the most important to begin is to be able to retrieve any of data in any blockchain and provide studies about these data. It's the role of explorer and some research made several use of these explorers [4], but we need to go further and get data directly from blockchain. Especially, with the help of good parsers we can imagine to be able to retrieve and represent data in a quite short time with making use of tools purposed in [26].

So the first step should be to make a deep study into Przytarski, Stach, Gritti et alt. [23] especially from the state of the art section and a review of the only serious paper in this field Deepa, Pham et alt. [6].
After this, we can try to purpose some experimentations of two types.

One with making use of AWS tools.
Another with the geth client and blockchain directly on the machine where are doing experiments. The second proposal was made in the master thesis of M. Kaandorp [13], and should be analysed with taking attention on the recent update of `geth` [8] client. Is it still interesting to do this ? Can we use this method and improve results with GraphQL native queries ?
A study about fog computing based on Sànchez, et alt. [24], could be done to improve our previous results by taking profit of decentralisation and cloud computing tools.

After this work, try an integration of tools deployed on European blockchain frameworks (EBSI) could be a good application of the previous work.

As there is more or less a lack of research in this way, it should be possible to produce new results and be published. Moreover, many companies could make several use of these tools and financing these research should definitely be possible.

5 Conclusion

Even though blockchain is fairly new [21] (2008), many things happened until now. Today, think about decentralised database is possible. Far to be effective now but possible. For many reasons using Bitcoin as a shared database is a really bad idea. But with tools develop by Europe (EBSI) it would be possible.

And think about new standards and new opportunities should obviously be done by knowing what is existing today !

Moreover experiencing actual data management on Blockchain should be a good way to go in this direction.

Appendix

EVM Schema

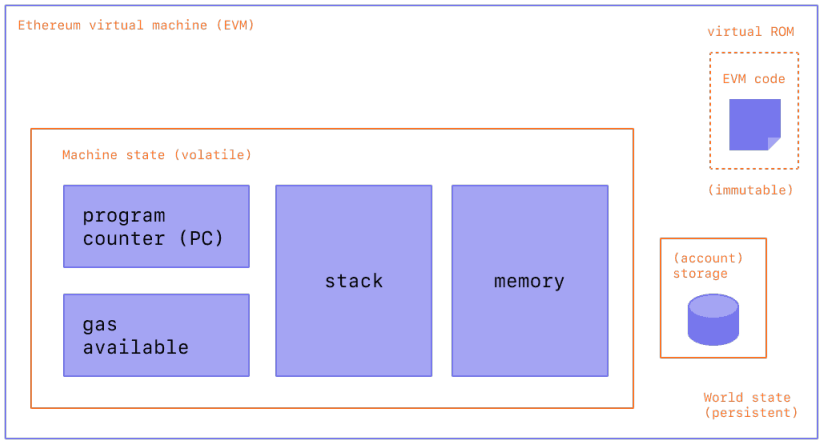

Figure 3: EVM Schema

Bitcoin recent update

As we know the blockchain in the case of Bitcoin is a Distributed Ledger Technology (DLT) [17]. It's mainly to store and transfer value all across the world with the only need of one address. But in a bitcoin transaction we can put a message.
With recent updates Segregated Witness (Consensus layer) (SegWit)[1] we can push much more data in a Bitcoin transaction. This update allows a several major change in Bitcoin protocol : the Taproot update [28][2].

Now there is a new protocol on top of Bitcoin to push data and retrieve it via an indexer. This protocol is called Ordinals. With making use of an enveloppe we can post and retrieve easily data.

This shown that we can build a basic database-like on top of Bitcoin. We can imagine how in a near future we would make use of Blockchain to allow many people

[1]This BIP defines a new structure called a "witness" that is committed to blocks separately from the transaction merkle tree. This structure contains data required to check transaction validity but not required to determine transaction effects. In particular, scripts and signatures are moved into this new structure. The witness is committed in a tree that is nested into the block's existing merkle root via the coinbase transaction for the purpose of making this BIP soft fork compatible. A future hard fork can place this tree in its own branch.

[2]This proposal aims to improve privacy, efficiency, and flexibility of Bitcoin's scripting capabilities without adding new security assumptions. Specifically, it seeks to minimize how much information about the spendability conditions of a transaction output is revealed on chain at creation or spending time and to add a number of upgrade mechanisms, while fixing a few minor but long-standing issues.

```
OP_FALSE
OP_IF
  OP_PUSH "ord"
  OP_PUSH 1
  OP_PUSH "text/plain;charset=utf-8"
  OP_PUSH 0
  OP_PUSH "Hello, world!"
OP_ENDIF
```

Figure 4: Ordinal enveloppe's

to store and retrieve data.

But it's a several change compare to our actual way to store data.

References

[1] Elvira Albert, Jesús Correas, Pablo Gordillo, Guillermo Román-Díez, and Albert Rubio. Gasol: Gas analysis and optimization for ethereum smart contracts. In Armin Biere and David Parker, editors, *Tools and Algorithms for the Construction and Analysis of Systems*, pages 118–125, Cham, 2020. Springer International Publishing.

[2] Vitalik Butherin. Ethereum: A next-generation smart contract and decentralized application platform. *Ethereum Foundation*, 2014.

[3] Pavel Ciaian, d'Artis Kancs, and Miroslava Rajcaniova. The economic dependency of bitcoin security. *Applied Economics*, 53(49):5738–5755, 10 2021.

[4] Wei Cui and Cunnian Gao. Wteye: On-chain wash trade detection and quantification for erc20 cryptocurrencies. *Blockchain: Research and Applications*, 4(1):100108, 2023.

[5] Weiqi Dai, Chunkai Dai, Kim-Kwang Raymond Choo, Changze Cui, Deiqing Zou, and Hai Jin. Sdte: A secure blockchain-based data trading ecosystem. *IEEE Transactions on Information Forensics and Security*, 15:725–737, 2020.

[6] N. Deepa, Quoc-Viet Pham, Dinh C. Nguyen, Sweta Bhattacharya, B. Prabadevi, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, Fang Fang, and Pubudu N. Pathirana. A survey on blockchain for big data: Approaches, opportunities, and future directions. *Future Generation Computer Systems*, 131:209–226, 2022.

[7] erh, dwight, and alt. Mongo db. Mongo DB | Github, 2007.

[8] Vitalik Butherin et alt. Ethreum client, go implementation. go-ethereum | Github.

[9] graphql. Graphql organisation. GraphQL | Github.

[10] Hossein Hassani, Xu Huang, and Emmanuel Silva. Big-crypto: Big data, blockchain and cryptocurrency. *Big Data and Cognitive Computing*, 2(4), 2018.

[11] Everett Hildenbrandt, Manasvi Saxena, Nishant Rodrigues, Xiaoran Zhu, Philip Daian, Dwight Guth, Brandon Moore, Daejun Park, Yi Zhang, Andrei Stefanescu, and Grigore Rosu. Kevm: A complete formal semantics of the ethereum virtual machine. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 204–217, July 2018.

[12] jbellis, pcmanus, and alt. cassandra. Cassandra DB | Github.

[13] Matthijs Kaandorp. Easy and efficient querying of smart contract data while maintaining data integrity. Master's thesis, Universiteit van Amsterdam, 2019.

[14] Aida Kamišalić, Renata Kramberger, and Iztok Fister. Synergy of blockchain technology and data mining techniques for anomaly detection. *Applied Sciences*, 11(17), 2021.

[15] Jan Kubal and Ladislav Kristoufek. Exploring the relationship between bitcoin price and network's hashrate within endogenous system. *International Review of Financial Analysis*, 84:102375, 2022.

[16] Sinan Küfeoğlu and Mahmut Özkuran. Bitcoin mining: A global review of energy and power demand. *Energy Research & Social Science*, 58:101273, 2019.

[17] Jennifer Li and Mohamad Kassem. Applications of distributed ledger technology (dlt) and blockchain-enabled smart contracts in construction. *Automation in Construction*, 132:103955, 2021.

[18] Jingming Li, Nianping Li, Jinqing Peng, Haijiao Cui, and Zhibin Wu. Energy consumption of cryptocurrency mining: A study of electricity consumption in mining cryptocurrencies. *Energy*, 168:160–168, 2019.

[19] Bin Liu, Xiao Liang Yu, Shiping Chen, Xiwei Xu, and Liming Zhu. Blockchain based data integrity service framework for iot data. In *2017 IEEE International Conference on Web Services (ICWS)*, pages 468–475, 2017.

[20] Paul Tak Shing Liu. Medical record system using blockchain, big data and tokenization. In Kwok-Yan Lam, Chi-Hung Chi, and Sihan Qing, editors, *Information and Communications Security*, pages 254–261, Cham, 2016. Springer International Publishing.

[21] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. `https://bitcoin.org/bitcoin.pdf`, oct 2008.

[22] Daejun Park, Yi Zhang, Manasvi Saxena, Philip Daian, and Grigore Roşu. A formal verification tool for ethereum vm bytecode. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2018, pages 912–915, New York, NY, USA, 2018. Association for Computing Machinery.

[23] Dennis Przytarski, Christoph Stach, Clémentine Gritti, and Bernhard Mitschang. Query processing in blockchain systems: Current state and future challenges. *Future Internet*, 14(1), 2022.

[24] Miguel Sánchez-de la Rosa, Carlos Núñez-Gómez, M. Blanca Caminero, and Carmen Carrión. Exploring the use of blockchain in resource-constrained fog computing environments. *Software: Practice and Experience*, 53(4):971–987, 2023.

[25] Shashank Tiwari. *Professional NoSQL*. Wrox, 2011.

[26] Natkamon Tovanich, Nicolas Heulot, Jean-Daniel Fekete, and Petra Isenberg. Visualization of blockchain data: A systematic review. *IEEE Transactions on Visualization and Computer Graphics*, 27(7):3135–3152, July 2021.

[27] Zhiyuan Wang, Zhiqiang (Eric) Zheng, Wei Jiang, and Shaojie Tang. Blockchain-enabled data sharing in supply chains: Model, operationalization, and tutorial. *Production and Operations Management*, 30(7):1965–1985, 2021.

[28] Pieter Wuille, Jonas Nick, and Anthony Towns. Taproot: Segwit version 1 spending rules. BIP 341 : Taproot, jan 2020.

[29] Meihua Xie, Haiyan Li, and Yuanjun Zhao. Blockchain financial investment based on deep learning network algorithm. *Journal of Computational and Applied Mathematics*, 372:112723, 2020.

[30] Li Zhang, Yongping Xie, Yang Zheng, Wei Xue, Xianrong Zheng, and Xiaobo Xu. The challenges and countermeasures of blockchain in finance and economics. *Systems Research and Behavioral Science*, 37(4):691–698, 2020.

[31] Weijia Zhang and Tej Anand. *Layer 2 and Ethereum 2*, pages 341–378. Apress, Berkeley, CA, 2022.

[32] Qiheng Zhou, Huawei Huang, Zibin Zheng, and Jing Bian. Solutions to scalability of blockchain: A survey. *IEEE Access*, 8:16440–16455, 2020.

[33] Jiawei Zhu, Peipeng Liu, and Longtao He. Mining information on bitcoin network data. In *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 999–1003, June 2017.