2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8292

Blockchain-backed analytics: Adding blockchain-based quality gates to data science projects

Herrmann, Markus; Petzold, Jörg and Bombatkar, Vivek
Technology & Data, GfK SE, Germany

Abstract

A typical analytical lifecycle in data science projects starts with the process of data generation and collection, continues with data preparation and pre-processing and heads towards project specific analytics, visualizations and presentations. In order to ensure high quality trusted analytics, every relevant step of the data-model-result linkage needs to meet certain quality standards that furthermore should be certified by trusted quality gate mechanisms.

We propose "blockchain-backed analytics", a scalable and easy-to-use generic approach to introduce quality gates to data science projects, backed by the immutable records of a blockchain. For that reason, data, models and results are stored as cryptographically hashed fingerprints with mutually linked transactions in a public blockchain database.

This approach enables stakeholders of data science projects to track and trace the linkage of data, applied models and modeling results without the need of trust validation of escrow systems or any other third party.

Keywords: *Blockchain; Data Science; Data Management; Trusted Data; Trusted Analytics.*

1. Trusted analytics

A typical analytical lifecycle in data science projects starts with the process of data creation and collection, continues with data preparation and pre-processing ("data wrangling") and heads towards project specific analytics, visualization and presentation, i.e. the results. To enforce trusted analytics, every step of the data lifecycle and applied analytics of an analytics project needs to meet certain quality standards. While these standards may vary broadly among academia and industries, there is one common challenge for every field of trusted analytics: How to publicly document trusted data and analytics in an immutable way? And if possible, without the involvement of any third party ensuring the trust.

Why is this considered a challenge? In academia, trusted analytics is achieved by peer-reviewed processes and bibliographical documentation. In data-driven industry sectors, on the other hand, the massive amount of decentralized data being generated daily and the huge number of data science and analytics projects worldwide cannot be evaluated and documented by any manual or human review system in a reasonable amount of time.

With the recently matured possibilities of machine learning – and in general the field of artificial intelligence - the documentation of the data-model-result relationship will become more and more relevant and consequently requires a scalable and immutable data and information documentation solution as a quality gate.

A decentralized storage system based on blockchain technology is able to introduce such quality gates to data science projects. For this reason we propose "blockchain-backed analytics"; whereby the data, applied methods and relevant results are stored as cryptographically hashed fingerprints in an immutable blockchain database.

Delivering this scalable and easy-to-use generic approach means being able to track and trace the linkage of data, models and modeling results, without the need of involving escrow systems or any other third party.

Our work builds on existing research in the fields of blockchain-based data protection and identity management, where blockchain technology is being applied to secure the management of digital identities and protect data ownership (Zyskind et. al, 2015). Accordingly blockchain-backed analytics is an extension of blockchain-based identity management techniques to data science projects and is therefore particularly relevant for data-driven academic research and industry projects.

2. Blockchain technology

To date the application of the blockchain technology is predominantly influenced by Satoshi Nakamoto's design of the cryptocurrency Bitcoin, which is based on the consensus in a distributed system, achieved with a Proof-of-Work algorithm (Nakamoto, 2008).

In this regard, a blockchain is a distributed database that is continuously keeping records of transactions in a logical order and in sync across participants, i.e. instances. Multiple transactions are bundled and stored in a block, whereas new blocks are sequentially appended to the previous block(s), with each block containing a list of cryptographically signed transactions with timestamps. In order to ensure integrity of the blockchain, each new block contains a pointer to a distinct hash value of the previous block and – in most cases – the root hash of the Merkle tree of all transactions of the previous block as well. This can be considered as a hash chain of all transactions of the block (ibid.).

The sequence of inter-linked blocks then forms a blockchain with the inherit feature that every block can be traced back to the initial first block of the chain. This also implies that any later modification or deletion of single transactions or entire blocks would result in a hash mismatch in hash pointers and Merkle trees and therefore break the chain.

A blockchain network can be private, where access and read/write permissions can be restricted, or public with unrestricted access and read/write permissions. Although the most popular applications of public blockchains to date are cryptocurrencies, the technology is by far not limited to this use case (Davidson, 2016).

The proposed public blockchain database approach seems to be most suitable for blockchain-backed analytics due to one of its core characteristics: the immutability of its records.

2.1. Immutability of the blockchain

The data stored in a blockchain database is immutable in the sense that once a record has been written, it cannot be modified or deleted afterwards. This can be put down to the process of validating transactions and adding them to a new block, which is commonly referred to as "mining".

Among others, there are two popular categories of mining algorithms in public blockchains: Proof-of-work (PoW), as applied with Bitcoin mining (Nakamoto, 2008) and Proof-of-stake (PoS), as proposed for the cryptocurrency Ethereum (Buterin, 2017).

These categories of mining algorithms both provide consensus among the distributed parties of the blockchain about the validity of the transactions and therefore, the final

commit to the database. Whereas each set of algorithms contains advocates and attack vectors[1], both sets also share characteristics which makes it almost impossible to determine the consensus process for the purpose of fraud or self-interest and therefrom derived ensure the immutability of existing blockchain records.

Considering that a broad distribution of mining instances is crucial to reduce the risk of manipulation, it is recommended to use a large (i.e. widespread distributed) public blockchain for blockchain-backed analytics. Alternatively a private or permissioned blockchain can be applied; in particular for big consortiums that aim to retain control over configuration parameters of the blockchain, e.g., to reduce transaction costs (Davidson, 2016).

2.2. Blockchain database capabilities

Despite its database structure, a distributed blockchain database is not primarily intended to be used as traditional database storage, mostly due to matters of the distributed technical design and mining process. Notably with the traditional Bitcoin blockchain, there are a number of known scalability limitations, such as the limited number of transactions per block, the limited throughput of transactions and the high latency until a transaction is confirmed (Croman et al., 2016). In addition, classical blockchains are usually not capable of any traditional querying capabilities (as opposed to RDBMS or NoSQL data stores) and in most cases only allow the lookup of existing – and thus valid – transactions.

For this reason, public blockchain databases mostly serve as distributed ledgers (especially for cryptocurrencies) with the property of providing synchronized, auditable and verifiable transactional data across multiple users and distributed networks (without the need of the involvement of third parties to validate transactions). They are not designed as data storage.

3. Blockchain-backed analytics

The idea of blockchain-backed analytics consists of creating an immutable linkage between the three core components of an analytics project: data, model and result.

The data component can be any kind of data that has either been used to train (i.e. to build) a model, or to apply a model. The model component can be any kind of data science model

[1] In order to manipulate the PoW consensus, the computing power (i.e. the hash rate) of a fraudulent participant needs to exceed 51% of the hash rate of the overall network. To fix a PoS consensus, a complex randomized process has to be determined. Hence, the probability of exactly hitting one of these attack vectors is negatively correlated with the size of the network and will tend to theoretically zero in large blockchain networks (Buterin, 2017).

represented as a function, script, library, binary executable, containerized application or even as virtual machine image; whereas the format of the result is determined by the model.

3.1 Blockchain signatures of components

Each component will be registered as a secure cryptographically hashed fingerprint as a transaction property to a public blockchain database, together with a pointer to the transaction identifier containing the component it continues from. The registration process consists of two steps:

1. Creation of the fingerprint (i.e. a secure cryptographical hash) of the component
2. Signature of the transaction to a public blockchain, consisting of:
    a. The hash of the component (**h_x**)
    b. The transaction identifier (**t_x**) of the linked component (optional for the data component)

The fingerprint should be created with a hash function in compliance with the Advanced Encryption Standard (AES)[2] and have a key length of no less than 128-Bit. The transaction properties can be submitted as a hexadecimal string, or ideally in JavaScript Object Notation (JSON) format (Fig.1.).

Figure 1: Transaction properties

```
{
 "properties":
  { "data": [{"name": "data ", "hash": "hx(data)"}],
    "model": [{"name": "model", "hash": " hx(model)", "data": "tx(data)"}],
    "result": [{"name": "result", "hash": " hx(result)", "model": "tx(model)"}]
  }
}
```

In short, this approach stores the linked chain of analytical components with an immutable public blockchain transaction, whereas each component can be always identified by its unique hash and transaction identifier. Records of the relationship of projects, hashes and transactions have to be kept separately.

3.2 Component linkage verification

In order to track a component, the blockchain can be queried by a given transaction or wallet identifier for a specific transaction that includes either data, model or result

[2] Advanced Encryption Standard (AES), Retrieved May 12th, 2018, from: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.197.pdf; (doi:10.6028/NIST.FIPS.197. 197).

information as transaction data. The data-model-result relationship can then be traced by the linkage of components that is being reflected in the result's transaction properties.

Furthermore, such a verification procedure could also simply be used to ensure the source integrity of data, models and results on an individual basis; by retrieving the component's signature and verifying the fingerprints against the fingerprints of the original or linked component. Notably, filtering queries (e.g. finding all datasets a specific model has been applied with) are in general not possible without parsing the entire blockchain.

3.3 Blockchain ecosystem

With the increasing popularity of cryptocurrencies, a vast set of blockchain implementations have emerged, with Bitcoin being the first in 2009. The blockchain technology best suited for a specific analytics project depends on individual requirements such as payload size, block time, transaction fees and the public availability of the blockchain. As a ledger for information verification, almost any blockchain technology that allows querying transactions and including transaction properties is principally applicable for blockchain-backed analytics.

Overall we can recommend the Ethereum blockchain as an ecosystem for blockchain-backed analytics. Component hashes can be either stored as raw transactional data (i.e. transaction property) in hexadecimal format, or alternatively integrated into a "smart contract", a programmatic feature of the Ethereum blockchain. Furthermore, as it is one of the largest public blockchain transaction networks worldwide, a widespread distribution of Ethereum nodes is guaranteed.

3.4 Costs analysis

Using a public blockchain network always involves costs to process a transaction, i.e. a fee must be paid before a transaction can be processed and validated.

With respect to the Ethereum ecosystem, the total fee for a single transaction adds up the base transaction price (currently 21000 "gas") and the costs for additional payload (currently 4 gas for a zero byte, 68 gas for a non-zero byte).[3] Considering that an AES 256-bit hash (equal to 32 bytes) can be expressed as a hexadecimal string with 64 characters, a complete data-model-result linkage documentation requires approximately 1 kilobyte of additional payload in hexadecimal format. In sum, the payload of all three components as additional hex-encoded raw transaction data of three transactions on the Ethereum

[3] Ethereum Homestead Documentation: Estimating transactions costs on the Ethereum blockchain, Retrieved May 12th, 2018, from: http://ethdocs.org/en/latest/contracts-and-transactions/account-types-gas-and-transactions.html.

blockchain resulted in about $0.20 total fees with a fast confirmation time (less than 30 seconds) in May 2018.[4]

In addition to transaction costs, blockchain-backed analytics involves computational costs for hashing the components. Since parallelizing the execution of computing a single hash is not possible, the computational costs of hashing a component only vary with the individual CPU performance; not with the number of physical CPUs or cores.

Our own performance tests with two popular secure cryptographic hashing algorithms (BLAKE2 & SHA-256) using commodity hardware have shown that large components with even a one terabyte file size can be hashed under 30 minutes and smaller components with sizes of up to one gigabyte within just a few seconds (Table 1.).

Table 1: Computational costs (time) for hashing different file sizes

CPU: Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz

Algorithm	1 MB	10 MB	100 MB	1 GB	10GB	100 GB	1 TB
BLAKE 2	0.004s	0.027s	0.181s	1.481s	15.584s	2m 25s	~25m
SHA-256	0.015s	0.110s	0.622s	6.204s	63.167s	10m 7s	~90m

Source: Own performance tests (2018).

4. Discussion

Whilst we believe that blockchain-backed analytics is a scalable and easy-to-use approach to ensure trusted analytics, there are several considerations which should be made.

For example, additional transaction costs for registering components in a public blockchain are not insignificant, although they are very low for single transactions. But it should be noted that – especially in the field of artificial intelligence – self-learning and self-evolving machine learning or deep learning models need to be tracked at every step of the model evolution process. However choosing the right blockchain technology (e.g. with optimal block size and transaction costs) for the specific project requirements and consolidating multiple components into a single transaction can help to optimize the costs of a blockchain-backed analytics project.

[4] Own registration of a result-component on the Ethereum blockchain on May 12th, 2018: https://etherscan.io/tx/0xd2749d1bcd7983769ba4801265c65fce8e92df7476f57df01bffcb148e5f0b32.

In theory, our approach is scalable to any kind of data size in terms of scalability and usability for big data applications, but in practice it is limited to the costs of hashing the components.

As described, the computational costs for hashing a component is not considered to be a significant overhead for component sizes up to a few gigabyte, but they have to be taken into account for big data. However in many big data environments data is mostly stored in distributed file systems, such as the Hadoop Distributed File System (HDFS), where the identification of distributed chunks of data is being achieved by inherit file checksum mechanisms that are already been applied during the data ingestion process.[5]

When applying blockchain-backed analytics with data or results stored in HDFS, the component does not need to be hashed again, because the available block checksums could be re-used as distinct block hashes in order to create a Merkle tree of all relevant blocks.

A similar approach also applies for containerized applications with inherit hashing mechanism, such as Docker images, where a fingerprint of the image is automatically created during the image build process.[6] Consequently, it is possible to easily integrate parts or entire analytical ecosystems in the format of a Docker image digest as distinct model component into blockchain-backed analytics.

Following our approach, where the data itself is not stored on the blockchain, an additional overhead process of maintaining a documentation of the relationship of projects, hashes and transactions has to be taken into account. However recent database solution developments with blockchain capabilities (e.g. decentralization and immutability) on top of traditional database capabilities (e.g. querying, indexing, search), could ease the adoption of blockchain-backed analytics, due to the omission of additional hashing procedures and documentation in off-chain references.[7] A similar ease of use could also apply for current developments of distributed (file) system solutions that are directly attached to a blockchain.[8]

---

[5] Hadoop Checksum, Retrieved May 12th, 2018, from: https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html#checksum.

[6] Docker Engine Reference, Docker images digests, Retrieved May 12th, 2018, from: https://docs.docker.com/engine/reference/commandline/images/#list-image-digests.

[7] e.g. solution "BigChainDB", Retrieved May 12th, 2018, from: https://www.bigchaindb.com.

[8] e.g. solution "The Interplanetary File System", Retrieved May 12th, 2018, from: https://ipfs.io.

With respect to continuous improvements in the development and integration of blockchain-based technologies, we are confident that our generic proposal of tracing the data-model-result linkage of analytical projects can be easily extended to broader ecosystems, such as continuous integration systems as part of the application lifecycle management.

References

Buterin, V. (2017). Proof of Stake FAQ., *Ethereum Wiki*., Retrieved May 12th, 2018, from: *https://github.com/ethereum/wiki/wiki/Proof-of-Stake-FAQ*.

Croman K. et al. (2016). On Scaling Decentralized Blockchains., *In: Clark J., Meiklejohn S., Ryan P., Wallach D., Brenner M., Rohloff K. (eds) Financial Cryptography and Data Security. FC 2016. Lecture Notes in Computer Science, vol 9604.* Springer, Berlin, Heidelberg.

Davidson, S., et al (2016). Economics of Blockchain., Retrieved May 12th, 2018 from: *https://dx.doi.org/10.2139/ssrn.2744751*.

Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System., Retrieved May 12th, 2018, from: *http://bitcoin.org/bitcoin.pdf*.

Zyskind, G., et al (2015). Decentralizing Privacy: Using Blockchain to Protect Personal Data., *2015 IEEE Security and Privacy Workshops*, San Jose, CA, pp. 180-184.