Blockchain-based transaction integrity in distributed big data marketplace

Denis Nasonov, Alexander A. Visheratin and Alexander Boukhanovsky

ITMO University, Saint Petersburg, Russia,
denis.nasonov@gmail.com

Abstract. Today Big Data occupies a crucial part of scientific research areas as well as in the business analysis of large companies. Each company tries to find the best way to make generated Big Data sets valuable and profitable. However, in most cases, companies have not enough opportunities and budget to solve this complex problem. On the other hand, there are companies (i.e., in insurance or banking) that can significantly improve their business organization by applying hidden knowledge extracted from such massive data. This situation leads to the necessity of building a platform for exchange, processing, and sale of collected Big Data sets. In this paper, we propose a distributed big data platform that implements digital data marketplace based on the blockchain mechanism for data transaction integrity.

Keywords: data marketplace, distributed systems, security, blockchain

1 Introduction

Today, it is impossible to imagine any area in business and no one branch of science, where large data could not be used to gain additional benefits or new knowledge. [10] However, for effective use of large volumes, it is not enough just to have them, it is necessary to understand how data can be used and how the results will give the desired benefit. Moreover, in many cases, large data has a higher value to related spheres of business rather its owner, where the obtained knowledge after processing the data can significantly affect the company's profit (for example, in the area of banking or insurance). This situation leads to the necessity of building a platform for the exchange, processing, and sale of collected big data. In this paper, we propose a distributed big data platform that implements digital data market, based on the blockchain mechanism for data transaction integrity. For the last nine years, blockchain-based technologies have attracted a lot of attention. Since the publication of Nakamoto's paper application areas of blockchain has expanded from cryptocurrencies to many other fields - databases, legal activities, medicine, etc. [1]. Researchers try to find novel ways to provide robust and scalable consensus [11] and to protect the data that is stored in blockchain [13]. Authors of [12] propose a second-layer, off-chain network Enigma that enables secure, decentralized data computation and exchange. One of the possible applications of Enigma is a data marketplace, but the idea

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_43

behind the proposed concept is closer to standard blockchain-based markets, where people sell their data, e.g., Datum platform (`https://datum.org`)[9]. In [4] authors provide a substantial investigation of existing projects related to data marketplaces and state that blockchain paradigm can be successfully applied to the development of data marketplace solutions. In another work, [7] authors proposed a market model for Big Data selling. However, they are concentrated near pricing optimization ideas without discussions on transaction integrity.

2 Data Marketplace concept

2.1 Problem statement

As big data marketplace is a new research area a lot of different questions appear in the field of interacting organization between customers and providers within dataset offerings in distributed data market platform. The main issues divide into four categories:

1. data processing integrity organization - formalized schema of used decentralized algorithms in cooperation between providers and platform, other providers and customers;
2. data processing limitations - discuss how Big Data should be processed in different interaction schema and which methods are allowed during processing (transfer, analysis, filtering, anonymization, simplification);
3. data validation procedure - specifies main principles of data checking that gives customer insurance in data veracity and truthfulness as well as checking for duplication offering;
4. data consolidation - defines methods how data from different resources can enrich processed data analysis and what can be done with anonymization.

All of this issues are crucial and should be investigated carefully in details. However, this paper is more concentrated on proposed data integrity approach and how it influences on other points.

2.2 Platform architecture

Basic platform architecture is presented in Figure 1. It consists of three main parts: DMP portal, DMP server and provider's infrastructure. Further, all the modules of the platform are briefly described for DMP processes understanding in the following section.

DMP portal includes digital data marketplace with showrooms and specific work environments. Environments are divided into three types: customer workspace, data holder workspace, and Administrative Tools. When new data holder joins the market, he registers and verifies his organization as well as create his data description, providing all necessary information. Data holder can also change already existed offers, including used policy, prices, and other conditions. Customer workspace allows the user to manage already bought data and

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_43

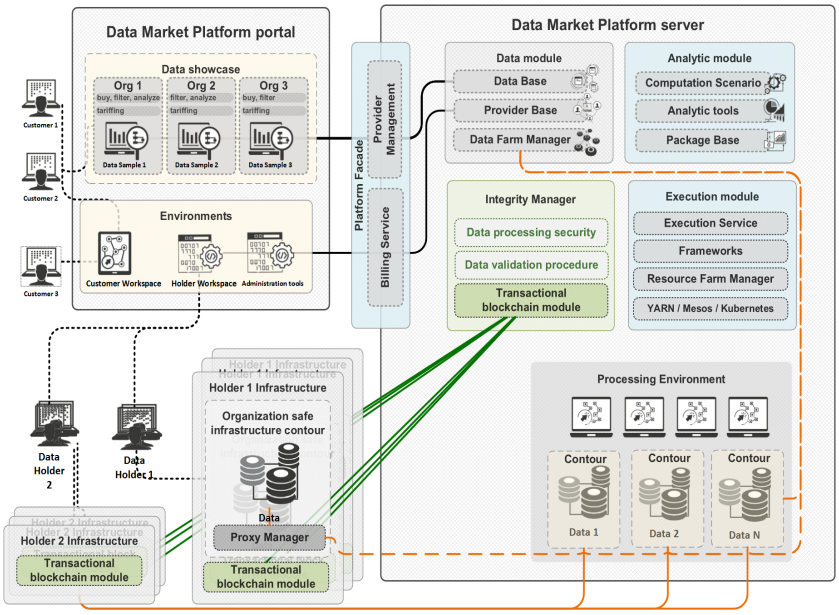

Fig. 1. Data Marketplace platform concept

provides access to remote data that can be only processed by selected methods in the data owner infrastructure. The user also may create different processing scenarios on workflow DSL using platform integrated analytic packages as well as other embedded tools. Administrative module hides all configurable system parameters, like user management, access policy, and showcase demonstration.

DMP server offers public functional API through the Platform Facade service, which extends CLAVIRE Facade Service [2] and manages user session activity. User session communicates with Billing service for financial requests and checks user rights for requested methods. However, the principle Platform Facade goal is to provide access to Data, Analytic and Execution modules.

The execution module is based on the CLAVIRE Execution Service [3] and combines several types of task execution on computational resources. The first type takes integrated frameworks (configured on Docker containers or VM), like Apache Spark or Apache Hadoop and processes data using provided users scenarios (in the form of workflow). The second type executes packages that were embedded in PackageBase (another CLAVIRE service) and were deployed on the available resources controlled by Resource Farm Manager.

An analytic module as Execution module is a part of CLAVIRE platform and contains services for packages and analytic tools management. Computational Scenario block interprets computational user requests extracting task parameters, which are described in Package Base and Analytic tools.

Data module has three blocks: Data Base, Provider Base, and Data Farm Manager. Data Base service includes a description of all providers' data sets that

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_43

currently are distributed through the market. It contains a description of meta information, such as data format, anonymization details, showcase conditions, etc. Data Farm Manager, in its turn, operates with all controllers of all registered in Data Base sets with enabled configurations and their locations. While Provider Base stores provider's meta information as well as offered datasets with limitation and applied policy. In other words, Provider Base defines all the rules upon methods and data access through the Analytic module and traditional selling.

In this paper, we concentrate our research studies on Integrity Manager module. It manages all essential transactions with data through its all life cycle, that guarantees data truthfulness and invariability for customers. The core of Integrity Manager is transactional blockchain module that is deployed on every registered provider and uses blockchain paradigm to make available trusted the decentralized system.

In data marketplace, there is a different option for data holders in data placement and data access. Data holder may decide to provide data access only within own private infrastructure contour that is remotely integrated with Data Farm Manager and Resource Farm Manager. Moreover, it can set strict access to specific analytical tools or even methods from Package Base. On the other hand, the provider may place his data in DMP infrastructure and provide user widespread access to the data processing tools. Finally, the provider can sell his data directly to the customer.

3 Data processing integrity

3.1 Data processing scenarios

In the data marketplace platform described in the previous section, one of the most crucial blocks is the Integrity manager, which is responsible for tracking internal operations over data, such as uploading of datasets by providers. The primary aim of this block is to make sure that the data used in processing operations is always the very same data, which was initially uploaded into the system. But since the marketplace itself has no direct access to the data, standard verification techniques are not applicable for this case, and there is a need for the development of the procedure, which would allow to control data integrity and prevent possible frauds by data providers.

There are two basic scenarios for data processing in the marketplace platform - data collecting and data analysis. The first scenario is quite simple - client searches for the data on the platform portal, orders the data and after successful payment, the data passes from the provider to the client. In the second scenario, the data is processed by platform's processing environment or provider's infrastructure depending on where the data locates. In both scenarios, it is essential to make sure that the data, on which the operations are performed, does not change over time and from client to client. A data processing security and validation procedure are used to improve confidence and security between data

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_43

marketplace participants. More detailed information about Integrity Manager and its components is presented it the following sections.

3.2 Blockchain-based solution

In this section, we describe the solution, which can provide integrity of the data in the distributed data marketplace platform. Because data providers do not have full trust to other providers and the platform, and to overcome the case of single point of failure, the integrity manager is decentralized by design and based on blockchain concept [6].

Integrity manager has two main components - data validation component and data blockchain instance. Data blockchain consists of blocks, each of which holds information about one dataset that was added to the platform. New blocks are generated when a provider adds new dataset in the system. Data can be added by uploading the data to the platform or by registering the data located in the provider's infrastructure. The process of block generation is depicted in Figure 2. Dataset is composed of a set of data slices, which are internally represented as byte arrays. Blockchain builds a Merkle tree [5] using data slices of the dataset. In this structure, the root node contains a hashed representation of all data slices, which cannot be decoded to obtain the initial data. After that, the root hash, the hash of the previous block, the unique ID of the provider and the list of key-value pairs $< ID, H >$ (where ID is a data slice identifier, and H is the hash of that data slice) are written to the data block, which is then added to the data blockchain.

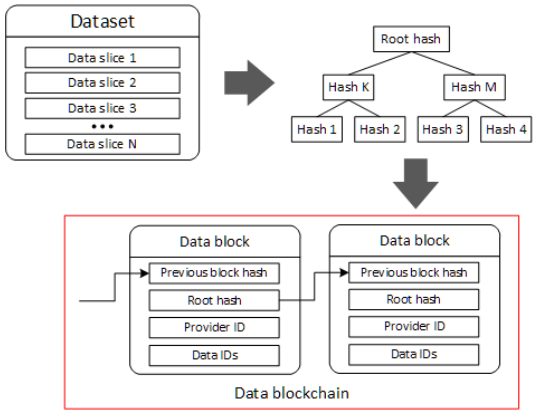

Fig. 2. Schema of blocks generation and chain structure.

Every provider, as well as the marketplace platform, holds its own instance of data blockchain. When provider adds a new dataset, the block is added first into its blockchain instance and after that is replicated to other instances in a

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_43

P2P manner. To prevent situations when two or more providers add their data at the same time, we apply simple consensus algorithm - if a collision of blocks is detected, the block, which was already synchronized across the largest part of the network, is held, other blocks are replaced by it. In this case, the data related to this block is discarded from the system and provider has to add the data again. This approach allows keeping the consistency between all instances of the blockchain. Although the described algorithm works good, there is a room for its improvement, e.g., usage of different consensus algorithms, such as Raft [8].

When a client tries to perform some operation on the data (collecting or analysis), this data is validated by all instances of the data blockchain in the following way:

1. Client requests an operation over some data slices to the marketplace.
2. Marketplace gets data slices from corresponding providers.
3. Data blockchain instance of the marketplace calculates hashes of data slices and validates them.
4. Data blockchain instance of the marketplace sends a validation request to all blockchain instances across the network.
5. Data blockchain instance of the marketplace collects responses from other instances until all responses come or more than 50% of all instances confirm that data slices are correct.
6. Operation requested by the client is performed over data slices, which were successfully verified.

As we can see, described method of data verification excludes a single control authority from the network, making possible reliable data verification without full trust between parties. The only considerable threat to this mechanism is collusion between data providers to change the data in the major part of the network. But since providers in the marketplace platform are usually big companies who take care of their reputation, such scenario is quite unlikely. Nevertheless, currently, we are working on a developed mechanism for elimination any possibility for blockchain interference. This mechanism is based on having a backup copy of the data blockchain, which has no collisions and is encrypted by every provider in the marketplace. Whenever any conflict in the network occurs, the marketplace would always be able to force the replacement of the current state of the blockchain by the backed up one.

3.3 Experimental study

To conduct an experimental investigation of the proposed blockchain-based approach, we have developed a simulator of the data marketplace platform, which implements logic described in the previous section. Source code for the simulator is available on Github - `https://github.com/visheratin/market-sim`. The simulator can perform following operations:

1. Create a specified set of data providers.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_43

2. Initialize data blockchain for the marketplace and all providers.
3. Add data for a specific provider and add a corresponding block into the blockchain. There is 0.001% probability that the uploaded data will be corrupted after uploading.
4. Search for the data according to user-specified criteria.
5. Collect data slices from providers and validate them through blockchain.

We conducted an experimental investigation of influence of data providers number and providers' reaction latency, which represents network latency and other overheads, to the speed of data verification. Providers number varied between 10, 100 and 1000; maximal latency was in range 10, 100, 1000, 5000 and 10000 ms. Search, and extraction was performed for correctly uploaded data slice and the corrupted one. Results of experiments are presented in Figure 3. We can see that with an increase in providers number the difference between validation time for correct and corrupted data equals up and goes closer to the average latency time. In the case of 10 providers, average validation time is much smaller than for other cases because in this scenario it is much easier to reach the consensus in data blockchain.

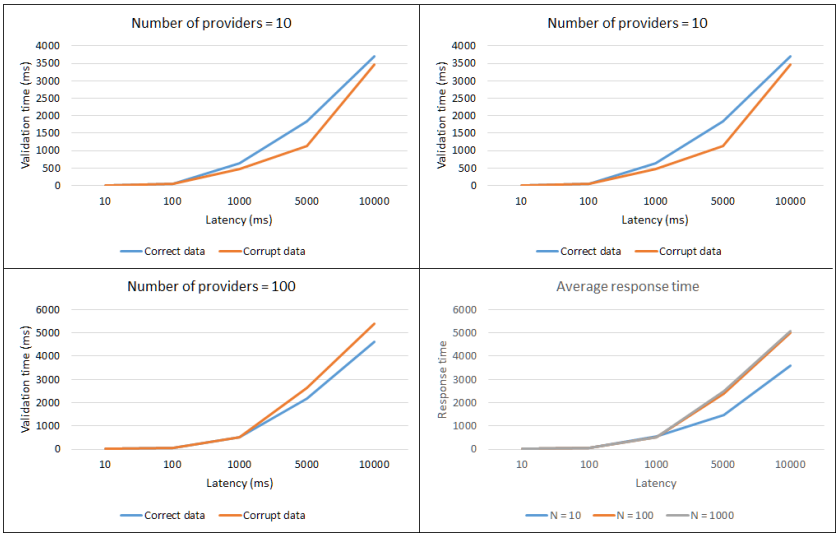

Fig. 3. Results of experimental evaluation of blockchain validation procedure.

4 Other features

To provide efficient data processing integrity between customers and provider we need: to guarantee that the data corresponds to the declared description

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_43

and the data could not be used in offering the second time on the DMP by another provider. To meet these issues, data validation, unification, and duplicates detection mechanisms are discussed below.

4.1 Data validation

Data validation helps the customer to be insured in data veracity and truthfulness. *Data prying* allows the user to check any shot part of the data in a randomized way for the limited amount of times. It helps the customer to make sure that showcase reflects the quality and characteristics of all data. For instance, customer intends to buy some measurement sensor data of level water in Baltic Sea; he can check the correctness and credibility of data in control points (some critical values on select dates).

4.2 Data consolidation procedure

Marketplace admits performing analysis on the enriched data from the aggregation of multiple sources (providers), including anonymous data. The algorithm is presented below:

1. Customer selects all dataset that should be merged;
2. Unique identification field is selected within all datasets (same UID in all data structures);
3. Upon each dataset, a hash function is applied to UID;
4. Data anonymization for all other field is performed;
5. All anonymized data sets are merged on platform infrastructure side using Data Consolidation Procedure;
6. UIDs of merged data sets are hashed one more time (in a case if some provider plays role of the customer);
7. Data is given to the customer.

This algorithm makes possible to combine open data (for instance, collected from social networks) with private data from financial or retail companies and conduct analysis over enriched data.

4.3 Duplicates, financial transactions

Duplicates can appear in two cases: provider tries to sell the same sets, or someone proposes already offered data. To avoid duplicates, the platform calculates statistical parameters as well as takes some randomized data markers. This information can be used by the customer to understand the difference between data sets and by providers to interpret the data originality. Integrated blockchain approach can also be used for financial interactions within the platform.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_43

5 Conclusion

In this paper blockchain transaction integrity within distributed Big Data marketplace concept was proposed. Experimental studies on developed simulator show appropriate results which inspire us to repeat them on a real system in the nearest future. Also different aspects of the platform, such as security, data truthfulness and union were discussed and proposed in present section above. However, we are looking forward to implementing all of them in a real system.

6 Acknowledgments

This work financially supported by Ministry of Education and Science of the Russian Federation, Agreement #14.575.21.0165 (26/09/2017). Unique Identification RFMEFI57517X0165.

References

1. Michael Crosby, Nachiappan, Pradhan Pattanayak, Sanjeev Verma, and Vignesh Kalyanaraman. Blockchain Technology - Beyond Bitcoin. *Berkley Engineering*, page 35, 2016.
2. Konstantin V. Knyazkov, Sergey V. Kovalchuk, Timofey N. Tchurov, Sergey V. Maryin, and Alexander V. Boukhanovsky. CLAVIRE: e-Science infrastructure for data-driven computing. *Journal of Computational Science*, 2012.
3. Konstantin V. Knyazkov, Denis A. Nasonov, Timofey N. Tchurov, and Alexander V. Boukhanovsky. Interactive workflow-based infrastructure for urgent computing. In *Procedia Computer Science*, 2013.
4. Pantelis Koutroumpis, Aija Leiponen, and Llewellyn Thomas. The (Unfulfilled) Potential of Data Marketplaces. *Working Paper*, 2420(53), 2017.
5. Ralph C. Merkle. A digital signature based on a conventional encryption function. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 293 LNCS, pages 369–378, 1988.
6. Satoshi Nakamoto. Bitcoin: A Peer-to-Peer Electronic Cash System. *Www.Bitcoin.Org*, page 9, 2008.
7. Dusit Niyato, Mohammad Abu Alsheikh, Ping Wang, Dong In Kim, and Zhu Han. Market Model and Optimal Pricing Scheme of Big Data and Internet of Things (IoT). 2016.
8. Diego Ongaro and John Ousterhout. In Search of an Understandable Consensus Algorithm. *Proceedings of USENIX ATC 14: 2014 USENIX Annual Technical Conference*, 22(2):305–320, 2014.
9. Xiaoqi Ren, Palma London, Juba Ziani, and Adam Wierman. Joint Data Purchasing and Data Placement in a Geo-Distributed Data Market. 2016.
10. Hai Wang, Zeshui Xu, Hamido Fujita, and Shousheng Liu. Towards felicitous decision making: An overview on challenges and trends of Big Data. *Information Sciences*, 2016.
11. Hiroki Watanabe, Shigeru Fujimura, Atsushi Nakadaira, Yasuhiko Miyazaki, Akihito Akutsu, and Jay Junichi Kishigami. Blockchain contract: A complete consensus using blockchain. In *2015 IEEE 4th Global Conference on Consumer Electronics, GCCE 2015*, pages 577–578, 2015.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_43

12. Guy Zyskind, Oz Nathan, and Alex Pentland. Enigma: Decentralized Computation Platform with Guaranteed Privacy. pages 1–14, 2015.
13. Guy Zyskind, Oz Nathan, and Alex Sandy Pentland. Decentralizing privacy: Using blockchain to protect personal data. In *Proceedings - 2015 IEEE Security and Privacy Workshops, SPW 2015*, pages 180–184, 2015.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_43