The blockchain as a backbone of GDPR compliant frameworks

Peyo HRISTOV[a]*, William DIMITROV[a]

[a]University of Library Studies and Information Technologies, "Tzarigradsko shose" Str.119, Sofia 1784, Bulgaria

Abstract

General data protection regulation (GDPR) is applied since 25 May 2018. It is designed to harmonize data privacy laws in European Union. It clearly defines what personal data is and identifies involved objects as Controller, Processor and Data subject. Although it focuses over the procedures in the organizations that have contact with personal data, it raises technological challenges about data storage, data processing, access control, identity management, system resilience, cybersecurity, post security breach measures, transactions traceability. We offer a new conceptual model with a trust management technology between controllers and data processors, which is based on the capabilities of the DLT. It can be useful in synthesizing software architectures, managing change due to the implementation of GDPR, ePrivacy, Policy Directive, and another forthcoming EU legislation. The article aims to present where the blockchain implementation can be helpful for the GDPR compliant operations. It doesn't cover the GDPR or blockchain in deep technical details, but just points out the important aspects where the DLT solution could be applicable. The paper is structured as follows: Section 1 introduces the paper. Section 2 describes the methodology used in this article. Section 3 introduces the GDPR and points out with short descriptions the key principles of the regulation. Section 4 covers brief explanation what blockchain is with examples from the Bitcoin implementation. Terms like transaction, transactions integrity, block, consensus are described here. Permissioned and permissionless blockchain implementations with their basic difference in the level of trust. Section 5 defines the main intersection points between GDPR compliance and the blockchain. Section 6 presents the Hyperledger fabric blockchain framework founded by Open Linux Foundation and IBM. The section focuses over the unique Hyperledger fabric abilities, which leads to increased confidentiality, transaction speed, traceability, access control, identity management, endorsement policies and smart contracts applications. Section 7 shows related research in this area. In conclusion the study reveals the biggest challenge in the blockchain application in the GDPR compliant frameworks.

Keywords: blockchain; confidentiability; GDPR; Hyperledger Fabric.

1. Introduction

General data protection regulation (GDPR) is applied since 25 May 2018. It is designed to harmonize data privacy laws in European Union. It clearly defines what personal data is and identifies involved objects as Controller, Processor and Data subject. Although it focuses over the procedures in the organizations that have contact with personal data, it raises technological challenges about data storage, data processing, access control, identity management, system resilience, cybersecurity, post security breach measures, transactions traceability. The blockchain became very popular since the end of 2017, because of the Bitcoin's price. But the blockchain as an implementation of the distributed ledger technology (DLT), has unique combination of some features like: traceability, transactions non-repudiation, very strong cryptographical nature, decentralization - which drew the attention of this study. In addition, smart contracts empower the blockchain with ability to automate transaction processing with programmed rules. The article aims to present where the blockchain implementation can be helpful for the GDPR compliant operations. It doesn't cover the GDPR or blockchain in deep technical details, but just points out the important aspects where the DLT solution could be applicable.

In this paper we offer a conceptual toolbox that solves three major problems:

- Strengthening the possibilities for complying with the principles of data processing and transmission and fulfilling accountability requirements;
- Operational mutual control over compliance with GDPR requirements in groups of related organizations;
- Provide opportunities for minimal impact of GDPR compliance on business continuity.

* Corresponding author. Tel.: +359-889-633-674
E-mail address: p.t.hristov@unibit.bg.

2. Methodology

The methodology of this article is based on a review over the essence of the GDPR and especially on the parts, where more complicated technical solution is needed. Bitcoin's blockchain is selected for the blockchain's overview, because it is quite simple and shows the most important aspects of the blockchain - transactions integrity, blocks sequence, consensus etc. Hyperledger Fabric official documentation and autor's experience in Hyperledger Fabric deployment and chaincode applications are used for explanation about multiple organizations (MSP) channels confidentiality and endorsement procedures, together with chaincode smart contracts. In the conclusion part the immutability vs. "Right to be forgotten" issue is raised as the main challenge in the GDPR and blockchain cooperation.

3. What a GDPR stands for?

GDPR is an abbreviation of General Data Protection Regulation. It replaces DPD (Data Protection Directive) 95/46/EC. It is designed to harmonize data privacy laws in EU to protect and empower all EU citizens data privacy and to reshape the way organizations across the region approach data privacy. In fact, GDPR is the most important change in data privacy regulation in 20 years (EU GDPR portal, 2018).

GDPR clearly identifies the following objects (ICO, 2018):

- Controller - determines the purposes and means of processing personal data.
- Processor - responsible for processing personal data on behalf of a controller.
- Data subject - a physical person. The personal data owner.

Definition of "personal data" and "processing" are also straightforward:

"'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;" (Concil of the European Union, 2016)

"'processing' means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;"(Concil of the European Union, 2016)

At a first glance, it seems that the controller is free from GDPR obligations, but this assumption is far away from the truth - the GDPR places further obligations on you to ensure your contracts with processors comply with the GDPR ((ICO, 2018)).

From what has been said above is obvious, that the huge part of the business is affected by this regulation. Gartner clearly says, that organizations are unprepared for GDPR(Gartner,2017). Unfortunately, Forrester also predicts, that 80% of the companies affected by the regulation, will not fully comply to it until 25th of May 2018 (Forrester, 2017).

The goal of this article is not to reveal in deep GDPR, but for completeness we should mention the basic principles of the regulation - they are its heart:

- Lawfulness, fairness and transparency
- Purpose limitation
- Data minimization
- Accuracy
- Storage limitation
- Integrity and confidentiality (security)
- Accountability

It is maybe not accidentally, that "Lawfulness, fairness and transparency" is placed on first. Staying away from the legal aspects of this principle, the fairness and transparency are the really important pieces here from the present article's point of view. However, we should remember that they are very close linked together, and violation of any of three fails the whole. Fairness could be violated for example, when the controller obtains personal data but the data subject is misled or deceived in this process. Hence, this clarifies the transparency requirement. The data controller must provide a clean and open (transparent) way, to can data subject approve the lawfulness and fairness pieces of the principle.

Purpose limitation. Purposes for collecting personal data should be clear. The organization is obligated to specify these purposes in a clear way in the privacy information to the individuals. The personal data could not be used for other purposes, except in the cases they are compatible with the original or the organization have a new consent from the customer.

Accuracy is another key topic in this research. According to the regulation, personal data shall be accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay. (Concil of the European Union, 2016)

Storage limitation says, that you must not keeping personal data than you need it. After that you must take measures to erase or anonymize it. Not only at controller's side, but also at all processors side. And of course, the organization must be able to proof that.

Integrity and confidentiality looks as the most related to the cybersecurity. But this principle isn't limited only to the way organization store and transfer data - organizational security is the real keystone here. It covers also how the data can be altered, accessed, deleted etc. Have the accessing entity enough permissions to do so and what access level exactly have and of course - who grants these rights. Have the organization working solution how to recover data after an accident and how will ensure its integrity between the parties. Security awareness education of the organization's staff takes a place here. The right risk assessment of the personal data stored, both with used hardware equipment, also have important role in this principle.

And at the end, here comes the Accountability principle. In short, the organization should be able to proof its GDPR compliance. Most noticeable in the scope of this article is the ability to detect, record and report security breaches. Data protection by design and more importantly - by default, is mandatory requirement here.

4. The blockchain

"A 'blockchain' is a particular type of data structure used in some distributed ledgers which stores and transmits data in packages called 'blocks' that are connected to each other in a digital 'chain'. Blockchains employ cryptographic and algorithmic methods to record and synchronize data across a network in an immutable manner." (Natarajan et al., 2017)

The blockchain main concept for linked blocks is more clearly illustrated on Figure 1.

Fig. 1 - Bitcoin's linked blocks headers.
hashPrevBlock - the hash of the previous block = (SHA256(SHA256(block header)))
hashMerkleRoot - the root hash of the merkle tree of all transactions in the block

Inside every block, are stored the transactions. The transaction in a Bitcoin's blockchain means a ledger for transferred amount of cryptocurrency from one owner to another, as shown on Figure 2.

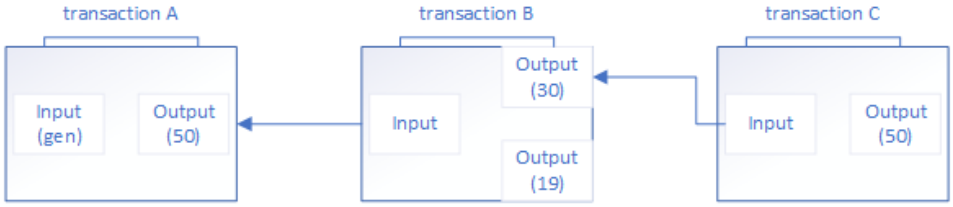

Fig. 2 - Transactions stored inside a Bitcoin's block. In transaction A, the owner initially owns 50 units. In transaction B, he sends 30 units to another owner in transaction C and keeps 19 units for himself. Note, one unit is lost for the transaction fee.

"Blockchains are distributed digital ledgers of cryptographically signed transactions that are grouped into blocks. Each block is cryptographically linked to the previous one after validation and undergoing a consensus decision. As new blocks are added, older blocks become more difficult to modify. New blocks are replicated across all copies of the ledger within the network, and any conflicts are resolved automatically using established rules."(Yaga et al., 2018)

As seen from Figure 1, every block (except the first block also called "genesis block") keeps the hash of the previous block header in hashPrevBlock header field. Noticeable here is hashMerkleRoot field, which keeps the integrity of the included in the block transactions. Hence hashPrevBlock relations between the blocks in the chain, protects the whole chain integrity. If an attacker modifies a transaction in earlier block, he will modify the hashMerkleRoot value which will affect attacked block header hash. After that, the next block's hashPrevBlock value will be wrong, so the attacker must alter all subsequent blocks in order to keep the blockchain's integrity. And here comes where the "Distributed" word in Distributed Ledger Technology takes place. Every new block or sequence of blocks (fork) should be successfully tested against some specific rules (for example longer fork wins) from a given amount of DLT nodes in order to be accepted. Moreover, every transaction must evaluate to true in a given application (also known as "smart contract") to become a part of a block. The whole process is called consensus.

According to the Hyperledger fabric documentation, consensus is a broader term overarching the entire transactional flow, which serves to generate an agreement on the order and to confirm the correctness of the set of transactions constituting a block. (Hyperledger, 2018)

The consensus is crucial part of the system based on DLT. That's because of the nature of different usage purposes. In the permissionless systems, there is an absence of trust. The consensus is based in a form of byzantine fault tolerance algorithm, based on "proof of work" which is based of huge waste of energy (Hern, 2017) and slows down the block acceptance speed and more practically important - the transaction speed.

But the enterprises or governments rare operates in environment with such lack of trust. They operate among a set of identified and known nodes. So, they operate in an environment with a certain degree of trust. Hence the permissioned model is more suitable for them and more simple and fast consensus protocols could be applied like crash fault tolerant and byzantine fault tolerant based. The DLT based on the permissioned consensus is in the focus of this article.

5. The points of interest

In a previous section, we take a brief look over the important concepts in GDPR. Specific organizational and technical measures should be taken in order to be compliant.

Transparency for example, is needed to the organization to can proof how the data is used. There should be a non-repudiation way to can organization proof who accessed this data.

Accuracy is another example. The data should be in-time altered or even erased if needed and all participating parties should receive the altered data or delete it.

Storage limitation is also complicated technical requirement. It also dictates, that data should be anonymized and deleted when no longer needed or at individual request. Also, reidentification risk should be taken in mind. Direct identifiers must be avoided - **"data that can be used to identify a person without additional information or with cross-linking through other information that is in the public domain."** (ISO, 2008)**.** Hence, the pseudonomization of personal data is a mandatory requirement.

Confidentiality and data integrity - how the data can be altered, accessed, deleted etc. Have the accessing entity enough permissions to do so and what access level exactly have and of course - who grants these rights.

Every data modification or access should be accounted. Data protection by design and more importantly - by default, is really required.

The requirements are so complex and of course single solution is hard to be found. At first glance, the DLT fits here. The blockchain implementation has its intrinsic abilities: data integrity, non-repudiation, data distribution. But at a second glance, some great issues can be found: lack of confidentiality, lack of access control, impossibility data to be forgotten etc. Yes, that is true. But only for the public (Zheng et al., 2017) blockchains like Bitcoin, Ethereum, Iota etc.

6. Hyperledger fabric

Hyperledger fabric framework (Hyperledger fabric, 2018) is one of the projects hosted by Linux Foundation together with IBM. With its modular architecture, it helps developers to create a blockchain based system for data distribution and control but avoiding permissionless blockchains drawbacks especially in the part of data confidentiality and access control. Thanks to its new architecture called execute-order-validate, the goals as flexibility, performance and confidentiality can be achieved in very convenient way.

Confidentiality comes together with a concept of channels. Channels are the virtual "way" to separate data available to organizations. Only allowed nodes could join the channel and it is nearly impossible rogue node to join in and eavesdrops the data in it. Node could be added in the channel only from an entity with the permission to do so. One node could join several channels. In the Figure 3 Organization 1 has two nodes in channel 1, but organization 2 has only one. However, Organization 1 peer 1 is joined in two channels channel 1 and channel 2, but channel 2 includes Organization 3 also.

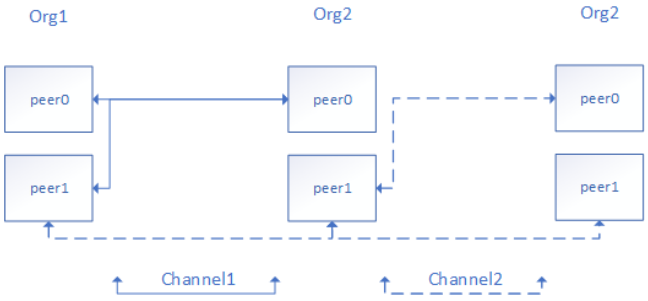

Fig. 3 - Hyperledger fabric channels concept. Three organizations with two channels. Transactions passed to the channel are distributed only over the participants in this channel. Every peer can join one or more channels (Org1.peer1), but they are still isolated from each other.

Node, that maintains the state of the ledger is called "peer". But the peer can handle a special role as endorser. Once the channel is created and the allowed peers are joined, the chaincode should be installed. Chaincode is an application written on Golang or Node.js, which defines the "smart contract".

Chaincode and the data in the channel are available only to the nodes, participating the channel, preserving confidentiality. Generally speaking, the chaincode "decides" if the proposed transaction is valid or no. In a greater detail, it could be very complex program with a lot of business logic inside.

Instantiation of the chaincode is important step afterwards. Instantiation must be completed on every peer, who should be an endorser. During instantiation endorsement policy is set. This policy defines which and how many peers (endorsers) should execute the installed chaincode on the selected channel against the transaction, in order to prepare it for future processing from the ordering service. The endorsement policy could look like:

"(Org1.peer0 OR Org1.peer1) AND (Org2.peer0)" which means the transaction must be endorsed from one of two Org1 peers and especially from Org2.peer0

If the received endorsements satisfy the policy, the client's node will send the transaction to the ordering service node also known as "orderer". Orderer "checks" the transaction endorsements again and if they are valid, "burns" the transaction into a block and broadcasts the block to the peers.

Hyperledger fabric provides identity management service that authenticates every network participant. Every step in the process flow is secured and satisfies "non-repudiation" principle.

From the GDPR's point of view it fits well technically in Accountability and Traceability requirements. It is undeniable who, when and what sent to the channel. It is also clear when the given organization is added in this channel and who grants the permissions to a given entity to do so.

Cybersecurity aspect in GDPR also could get benefits. Exceptionally strong cryptographic nature of the blockchain combined with identity management provided by Hyperledger fabric, makes the data stored in the blockchain resistant from unauthorized modification. In his article focused over blockchain's role in a cybersecurity (Kshetri, N., 2017), Nir Kshetri makes a deeper research about blockchain's role in strengthening cybersecurity and data privacy. He also noticed the blockchain's role in a post-breach resilience plan and implementation.

The technology has also another great benefit - it is distributed. That means, that even a large set of peers to be out of order, if the topology is scaled properly the availability of the service is not affected. When the failed peer goes in active state, it could sync the blockchains from any other peer.

Hyperledger fabric is applicable in the Accuracy also. Hyperledger fabric distributed ledger consist of two components. World state and transaction log. World state keeps the current ledger state, but transaction log keeps all previous transactions which leads to the current world state. Thus, once the transaction is ordered and broadcasted to the peers, all parties included in the channel will see the new data almost immediately.

7. Related work

The possible blockchain usage together with GDPR frameworks isn't isolated research. An elegant solution for "The right to be forgotten" conflict with the blockchain's Immutability, proposed the Chainfrog company in its article (Chainfrog, 2017). They propose several options like: record personal data pseudo-anonymously; encrypt the data on the blockchain; store the data in a referenced encrypted database. But Chainfrog's article raises another concern, about the right of EU persons to contest automated decisions which addresses smart contracts - as programmed logic aka. automated decisions. Another article that analyzes the impact of the GDPR over the blockchain technology is "A guide of blockchain and data protection" (Maxwell and Salmon, 2017). It provides a good overview about hashing technology, variety of blockchain systems, blockchain data protection impact assessments and their impact with data protection principles. In parallel "Gran Thornton" company (Gran Thornton, 2018) also provides solution about data removal, data modification and confidentiality, based on logical "private channels" with encrypted data available only to the participating nodes. Each time when nodes need to update or remove data, the corresponding decryption key of the old data should be removed.

8. Conclusion

The blockchain framework Hyperledger fabric has a lot of advantages, which can be used in conjunction with any software for data processing, even those required for very strict GDPR operations. Hyperledger fabric scalability, modularity, all levels identity management, confidentiality and privacy features build on top of intrinsic blockchain distributed and secure nature, seems to be the right technology candidate here. The benefits from such framework architecture would be:

- It facilitates the fulfillment of contractual obligations of controllers and processors involved in the transmission and processing of personal data;
- It strengthens the credibility of participating organization's confidentiality policies;
- It integrates a level of trust in data exchange interfaces with other organizations;
- Ensures evidence in inspections and investigations;
- Creates opportunities for clear policies and procedures for personal data breaches.

But there is an issue. An issue which confirms the sentence "there is no single development, in either technology or management technique, which by itself promises even one order-of-magnitude improvement within a decade in productivity, in reliability, in simplicity". (Brooks, F., 1986) The issue with a fundamental blockchain feature: The blockchain is immutable.

It is very difficult to do any modification in a block already added. That could lead to some difficulties from the GDPR point of view. Especially with one right of the data subject - Right to erasure ("Right to be forgotten"). Even if the controller issues a transaction to delete data from the world state component, the data is still in the transaction log. It is still in the blockchain, stored securely in every one peer's storage.

The present article leaves this problem for a future research. The possible solutions should not be searched with blockchain hacks like "editable blockchain" for example. The better approach should be to combine existing technologies patterns with the newly unleashed features of the blockchain.

Acknowledgements

This research was supported by the Bulgarian FNI fund through the project "Conceptual Modeling and Simulation of Internet of Things Ecosystems (KoMEIN)", contract DN 02/1.

References

Brooks, F. P., No Silver Bullet-Essence and Accident in Software Engineering, *Proceedings of the IFIP Tenth World Computing Conference*, (1986).

Gran Thornton, *GDPR & blockchain - Blockchain solution to General Data Protection Regulation*,Gran Thornton,(2018).

Hern, A., *Bitcoin mining consumes more electricity a year than Ireland*, *The Guardian*, (2017).

Kshetri, N., *Blockchain's roles in strenghtening cybersecurity and protecting privacy*. Telecommunications policy, (2017).

Maxwell W., Salmon J., *A guide to blockchain and data protection*, Hogan Lovells, (2017).

Natarajan H., Krause S., Gradstein H.,Distributed Ledger Technology (DLT) and Blockchain. *http://documents.worldbank.org/curated/en/177911513714062215/Distributed-Ledger-Technology-DLT-and-blockchain*, (2017).

Yaga D., Mell P., Roby N., *Scarfone K.,Blockchain Technology Overview*. Draft NISTIR 8202, (2018).

Zheng., Xie S., Dai H., Chen X., Wang H., An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends, *2017 IEEE 6th International Congress on Big Data*, (2017).

Concil of the European Union., *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC* (General Data Protection Regulation), http://data.consilium.europa.eu/doc/document/ST-5419-2016-INIT/en/pdf, (2016).

Chainfrog, *Blockchain and GDPR*, Chainfrog, (2017).

EU GDPR portal., *https://www.eugdpr.org/*, (2018).

Forrester., Predictions 2018 A year of reckoning, *Annual edition predictions*, (2017)

Gartner., Gartner Says Organizations Are Unprepared for the 2018 European Data Protection Regulation, *Gartner newsroom*, (2017).

Hyperledger, *hyperledger-fabric docs Documentation Release master*. (2018).

Hyperledger fabric, <*https://www.hyperledger.org/projects/fabric*>, (2018).

ICO., Guide to the General Data Protection Regulation (GDPR), *https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/*, (2018).

ISO., *Health informatics*. ISO/TS 25237:2008, (2008).