Towards Trustworthy and Independent Data Marketplaces

Priyanka Sharma
Institute for Software and Systems Engineering, Clausthal University of Technology
Arnold-Sommerfeld-Straße 1
38678 Clausthal-Zellerfeld, Germany
+49 5323 72-7193
priyanka.sharma@tu-clausthal.de

Sebastian Lawrenz
Institute for Software and Systems Engineering, Clausthal University of Technology
Arnold-Sommerfeld-Straße 1
38678 Clausthal-Zellerfeld, Germany
+49 5323 72-7176
seabstian.lawrenz@tu-clausthal.de

Andreas Rausch
Institute for Software and Systems Engineering, Clausthal University of Technology
Arnold-Sommerfeld-Straße 1
38678 Clausthal-Zellerfeld, Germany
+49 5323 72-8232
andreas.rausch@tu-clausthal.de

ABSTRACT
Data is the new oil. In the past years the awareness about benefits of data has increased. A growing number of sectors recognize the opportunities from data. On the one hand it is very difficult for many researchers and enterprises to obtain data and on the other hand for those who collect data, the problem is, how to draw to additional profit from the data beyond its obvious purpose. To tackle this problem, we propose a common data sharing platform, where the data producers can sell data and the others can consume it. Just like any other online marketplace a data marketplace is a platform which enables convenient buying and selling of products- in this case "data". Blockchain enables businesses to be decentralized and more secure. Thus, in this paper we explore an approach to combine data marketplaces and blockchain for fair and independent data marketplaces.

CCS Concepts
• Information systems → World Wide Web → Web applications → Electronic commerce → Electronic data interchange • Security and privacy → Systems security → Distributed systems security • Computer systems organization → Architectures → Distributed architectures

Keywords
Data marketplaces, Data economy, Blockchain, Smart contracts, Data Quality, Security and Privacy

1. MOTIVATION AND INTRODUCTION
Data today plays a much bigger role than it used to and now data is the new oil [1].

On one hand it is very difficult for enterprises to obtain proper data and on other hand for those who collect data, the problem is- *How to draw to additional profit from the data beyond its obvious purpose*. A data marketplace is required to facilitate trade and exchange of data and money. A data marketplace would serve as a trading platform where the data producers can sell the data and others can buy it. Thereby, data marketplaces will foster innovative business models where data is the raw material – data becomes the primary commodity. However, how to trade this commodity is still a challenge.

ICBCT'20, March 12–14, 2020, Hilo, HI, USA

ACM ISBN 978-1-4503-7767-6/20/03…$15.00
https://doi.org/10.1145/3390566.3391662

And this although nowadays a bunch of various electronic marketplaces – so called electronic commerce platforms – already exist, like for instance eBay, Amazon or Alibaba. These electronic marketplaces are platforms or infrastructure that allows participants to meet and perform their desired market transactions through an electronic medium. Thus, a data marketplace can also be categorized as a form of electronic marketplace. Electronic marketplaces exist in different shapes and can be categorized along various dimensions. As a result of the overlapping definitions of electronic marketplaces, the categorizations are equally confusing. Each model uses different definitions which makes a general classification of the various forms of business models difficult [2].

In [2] a comprehensive model incorporating various dimensions for categorizing electronic marketplaces is proposed. For the categorization of electronic marketplaces in our work we consider this model.

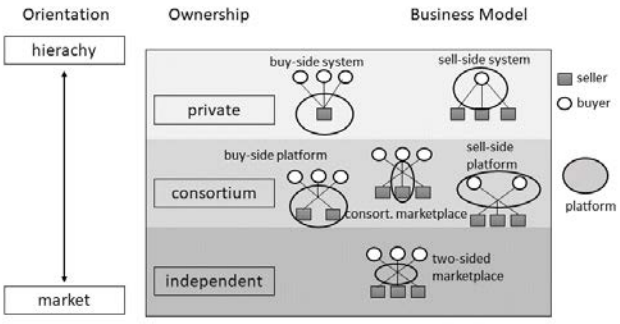

Figure 1. A model of electronic marketplaces discerning between three ownership types. Based on [2]

In this model first, providers are placed on a scale of orientation between hierarchy and market. Economies have two basic mechanisms for coordinating the flow of materials or services through adjacent steps in the chain: markets and hierarchies. Markets coordinate the flow through supply and demand forces and external transactions between different individuals and companies. Market forces determine the design, price, quantity, and target delivery schedule for a given product that will serve as an input into another process. Hierarchies, on the other hand, coordinate the flow of materials through adjacent steps by

controlling and directing it at a higher level in the managerial hierarchy, rather than by letting market transactions coordinate it. Managerial decisions, determine design, price, quantity, and delivery schedules at which products from one step on the chain. Thus, all transactions between suppliers and buyers can be classified as either hierarchical or market-based. Furthermore, marketplaces are categorized in based on their ownership, which can be (a) private, i.e., owned by a single company (seller or buyer); (b) consortia-based, i.e., owned by a small number of companies (seller or buyer); and (c) independent, i.e., the marketplace is run as a platform without any connection to sellers or buyers [2].

Based on these three dimensions market, hierarchy and ownership the model proposed by [2]differentiates six business models. As shown in the fig. 1 at the top are the privately-owned platforms. These types of business models typically facilitate the selling and buying of its business owners i.e. a company and only allows one-to- many and many-to-one relations. In between are the consortia- based platforms, these models are typically a collaboration between various companies to facilitate their buying and selling methods. At the market-level many-to-many marketplaces are usually operated by the independent parities and have minimal entry restrictions [2].

From this classification of electronic marketplaces presented by [2] it can be seen that the facilitation of procurement and selling in privately owned or consortia-based platforms i.e. owned by a group is towards themselves. An independent marketplace at market-level without any bias towards the buyers or sellers would serve as neutral platform [2]. In order to facilitate fair and un-bias trade more independent marketplaces are required. Thus, to facilitate a fair trading and independent marketplaces should be secure and not facilitating procurement and sales of the owners but allowing sellers and buyers to sell freely in a many-to-many relationship.

As data marketplaces are also a type of electronic marketplaces in order to establish an un-bias and independent data trading platform the data marketplace should be secure and facilitating the users fair and secure trade. But currently there are not many independent marketplaces existing which lets a user sell or buy quality datasets. There are many reasons for this gap, but we identify that most of the clients trust the private parties for providing reliable services, security and privacy. In order to have an independent buyer to seller relationship without any bias towards buyers and sellers, trust in the system is missing just like the consumers have trust in the middleman to provide them with the services. Therefore, in order to have an independent buyer to seller data-marketplace a trustworthy, secure, efficient platform in required.

In recent years a new technology called blockchain evolved which has the potential to provide a trustworthy, secure platforms for peer to peer transactions. Blockchain could be seen as a public ledger and all committed transactions are stored in a list of blocks. This chain grows as new blocks are added to it continuously. Cryptography and distributed consensus algorithms are implemented for securing and keeping the ledger consistent [3] . Blockchain first caught attention through bitcoin which enabled users to send money to each other directly without a need of financial institution [4]. But bitcoin is just one example of how the underlying technology blockchain can enable peer-to-peer trading. We identify that blockchain can enable more independent trustworthy and secure data-marketplaces without any bias towards the buyers or sellers. Thus, in this paper we are tackling the following research question:

How can an independent and trustworthy data marketplace leading to a new data economy could be implemented using the blockchain technology?

In our last paper “Blockchain technology as an approach for data marketplaces ” we outlined already how blockchain can be a underlying technology of data marketplaces and the associated challenges with them [5]. This paper gives more insights into some solutions and open challenges.

The rest of the paper is structured as follows section 2 gives a brief overview of related work. The requirements and interaction of the main stakeholders with the data marketplace is shown is section 3. An approach for the identified challenges is presented in section 4. Section 5 outlines open challenges. Finally, section 6 concludes and gives insights of future work

2. STATE OF THE ART

Real-time data will increasingly turn into a commodity in the coming years. With the aim of providing real-time data, Streamr is a data marketplace market for real-time data. In the data market, anyone can publish events to data streams, and anyone can subscribe to streams and use the data in decentralized apps. Most of the data is free, and the terms of use are stored in Ethereum smart contracts [6]. In the recent years there has been many ideas proposed with the aim of giving back the data ownership to the data producers and let them decide whether they want to share their personal data or not. One such project is datum; the Datum Client empowers users to take control of all their data and optionally share or sell their data through the Datum network [7]. The IOTA Marketplace[1] is a decentralized data market place that aims to make IOT data available to any compensating party. The Mobility Data Marketplace[2] (MDM) enables different parties to offer mobility data, such as petrol prices, or construction sites on motorways. Whereas projects such as streamer and datum are using blockchain to enable real-time data streaming and empowering users to take control of their data our goal is to establish a data marketplace where users can sell and buy quality data like any other commodity sold electronically today. Other platforms like *Kaggle*[3] from Google offer datasets for free.

3. REQUIREMENTS

Before delving in the architecture of our approach, we describe the requirements of the system that we want to satisfy:

There are some functionalities of a system like a data marketplace which are required for the data marketplace to always function. We define these requirements as functional requirements independent for the type of platform. But during our research there are some other requirements that we identify for data trading and for a fair and open platform. These requirements are described as non-functional requirements.

3.1 Functional Requirements

Fig 2. Shows the functional requirements of the data marketplace

[1] https://data.iota.org/

[2] https://www.mdm-portal.de/

[3] https://www.kaggle.com/

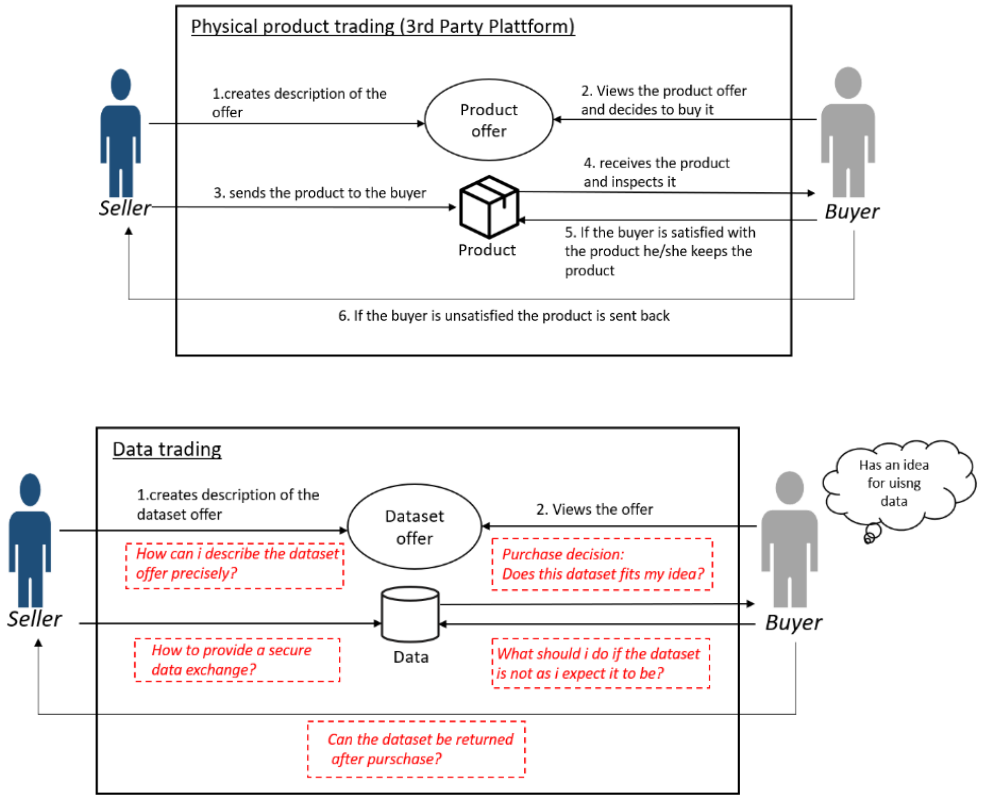

Figure 2. Physical product trading vs data trading

1. The system should allow sellers to add a sales posting and buyers to search for data: The main purpose of a data marketplace is to bring interested parties together and to facilitate users to sell and buy data, thus the system should allow buyers to add their selling post. The marketplace should also index all the postings and allow buyers to find streams of data interesting to them.

2. The system should allow the parties to trade data: Apart from adding a selling post and finding relevant data, the marketplace's main function should be in sharing the selected data with the right person.

3. The system should support payments methods: Generating funds by selling data is at the core of the data marketplaces business model, so the system must have payment methods established for the trade.

3.2 Non-functional Requirements

As already mentioned in the introduction the main problem towards establishing a data economy is the lack of trust. The lack of trust is sometimes in the involved parties, but also in the system itself. In many cases the data provider and the data consumer can't trust each other easily, because they do not know each other or have never met. Apart from the missing trust in the involved parties, trust the overall system (here the data marketplace) is also missing. Thus, the system has to guarantee trust in a way that the involved parties can trust each other and also trust the system while using it.

1. *The architecture of the system should be as open and transparent as possible:* Our motivation is to establish a fair, unbiased and transparent data exchange platform where the parties do not need to trust a middleman but can have trust in a system.

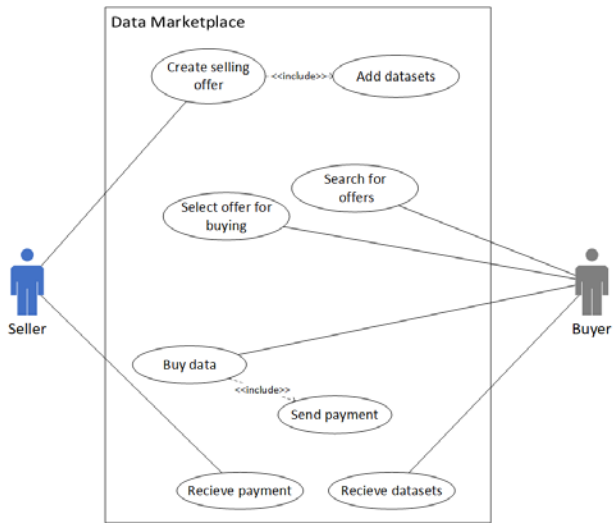

Figure 3. Functional requirements of a data marketplace

2. The system should allow data trading: Data as a commodity, seems similar to any other product sold online. But data is quite different than other products sold online. Fig 3. shows how data is trading is different then and other product sold online. Although there are many differences, we identify the following as the two biggest differences between other products and data i.e., Product description and product quality [8].

Product description: When any product is sold online, the product can be described in many ways. In many cases a detailed product description and with the help of pictures as well as videos can be provided. This helps the buyers a lot in the decision-making process. But in case of data describing the datasets for selling precisely is a big challenge. The additional details such as pictures and videos cannot be provided for data thus the data description (i.e., what is the data about, where was it collected etc.) is the most important factor affecting the buyers decision [8]. This additional information about the datasets or information about the data is known as metadata. To give all the stakeholders of a data marketplace an overview about the content and offers, metadata are essential.

2.1. The system should generate metadata automatically: As metadata is necessary to identify relevant data in a data marketplace, in order to help the seller to create an offer, to give the buyer a first overview about the data set and much more. Metadata is one of the most important factors for a data marketplace In many cases the seller might not be able to describe the offer properly, or the seller might not give a correct description of the datasets. Thus, we identify that the system should be able to generate metadata.

Product quality: Most of the products sold online except for digital products such as movies, songs etc. can be returned if the customer is not satisfied with the product. But the same cannot be done with data. As once the data is seen by the buyer it loses its value and the buyer might make copy of it. Thus, return policies for data is not feasible [8].

2.2. The system should be able to check data quality without storing the datasets:

Metadata serves as description of the data but does not describes the quality of datasets. We divide the quality of data in two criteria's: objective criteria and subjective criteria [5]. The objective qualities of data such as density, actuality etc. can be defined by metadata. But the subjective properties are the relevancy of the data with regards to the buyers' requirements. Thus, the quality of the data needs to be checked according to the buyer's specific requirements. But the buyer's requirements need to be hidden from the seller. As once the specific requirements are known fake data can be generated. Also, many users look for some specific data sets for their businesses. And showing the requirements can reveal their business models. For example- A buyer has a business idea for predicting state of the health of traction batteries using machine learning. For this he/she needs a various feature's in a dataset. When he/she checks the datasets for these requirements, the business logic of traction battery health prediction could be revealed. Thus, we realize that for data trading and establishing trust in system is necessary. It should be possible that the system can check the quality of data without storing the actual data and without revealing any business logic.

3. The system should be flexible: One of the common seen factors in platforms and many other data marketplaces is that they are not flexible. We realize that many users have different requirements in terms of data storage and payment options. Rather than forcing the users to store data and buy data from one particular storage option, in our approach, the user would be given multiple storage and payment options thus making the system as flexible and user friendly as possible.

4. The system should be secure and protect privacy of the users: Security is one of the most important requirements of any system, and it's usually very difficult to achieve. But when approaching towards independent and transparent marketplaces, security becomes a very critical requirement. The system should establish secure data trading and transactions. It should also validate the transactions and if the correct data is transferred to the rightful owner.

4. ARCHITECTURE

The overall architecture of our proposed system is outlined in fig.4.

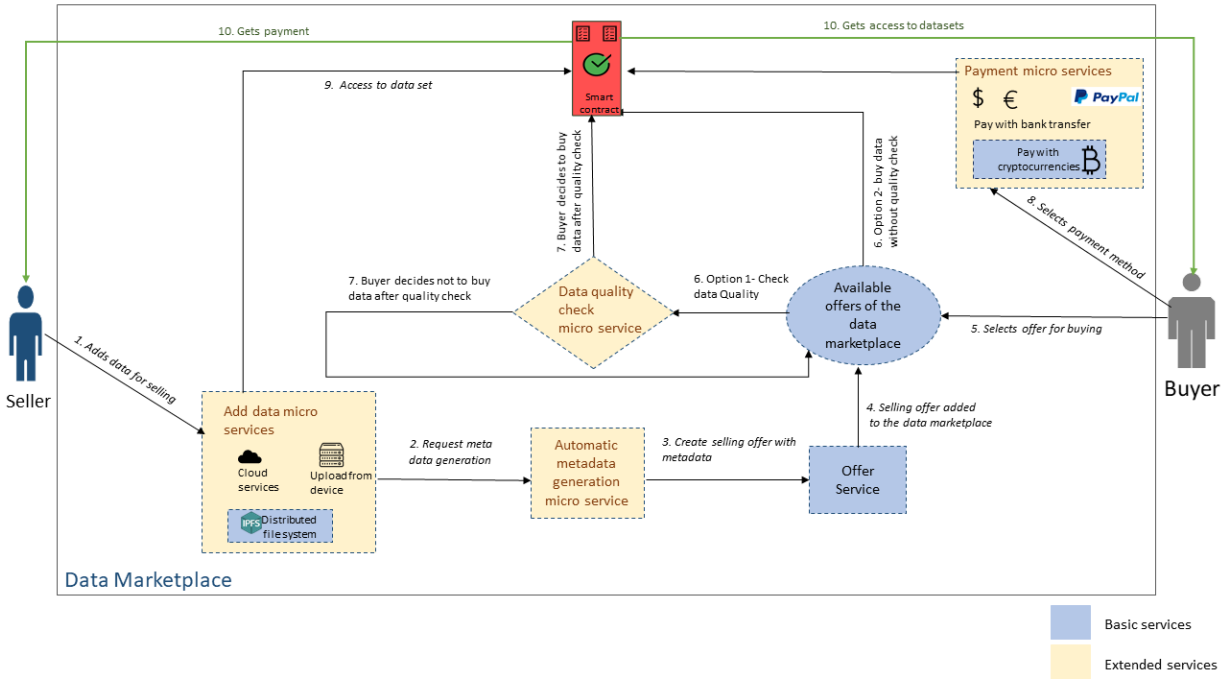

Figure 4. Overall approach

Our proposed fair and independent data marketplace is based on a microservice architecture. Microservices are independently deployable services modeled around a business domain. They communicate with each other via networks, making them a form of distributed system. They also encapsulate data storage and retrieval, exposing data, via well-defined interfaces. So, the data bases are hidden inside the service boundary [9].

Rather than increasing the system's complexity and building a single huge application to fulfill various requirements, we decided to adapt to a microservice architecture for various reasons. But the main reason for considering it was our two different requirements sets i.e. Functional requirements and Non-functional requirements. Having such a system architecture would allow the system to always function with the basic functionalities and still would allow constant development of the extended requirements without disturbing the whole system.

Microservices would allow decentralized data management, modularization and services with strong boundaries and technology diversity.

Based on the functional and non-functional requirements the architecture is defined with two types of services: basic services ( ) and extended services ( ). The functional requirements are the basic services and the extended services fulfill the non-functional requirements.

Basic Services: The basic services would allow trading of data without any extended functionalities. Although the data marketplace running only basic services would lack some special requirements, but it would still allow peer to peer data trade.

1. Seller wants to sell data using the data marketplace, and thus needs to create an offer and also has to add the datasets. By default, as a basic service currently the user can upload the datasets from his/her device to IPFS[4]. The Interplanetary File System (IPFS) is a protocol and peer-to-peer network for storing and sharing data in a distributed file system. After the files is uploaded a unique hash is generated.

2. After uploading the datasets the seller can add some description of the datasets and the offer is added to the data marketplace for selling.

3. The buyer on the other side can look through all the available datasets offer and can select a relevant dataset.

4. Once the buyer selects the dataset, its added to the shopping cart and the buyer is directed to the payment service.

5. If the payment is valid, the payment details enter to a smart contract. The smart contract holds the hash of the selected dataset and the payment. After all the validation the buyer gets the access to the dataset and the seller gets the expected amount for the datasets. If the payment or the hash is invalid, the transaction is aborted. All the successful transactions are published on blockchain for future validation.

The smart contract here is the secure middleman, which would always validate all the transactions. The storage of data in IPFS gives the control of the data to the users and no control of the data to the system.

[4] https://ipfs.io/

Extended Services: All the extended services are derived from the non-functional requirements. The extended services are the core of our approach.

1. Multiple storage options: The add data micro services has various data storage options. Due to this feature, the buyer doesn't need to place the data from his/her storage to another location, but can directly sell the data from wherever he/she has it. For e.g. if the data is in a cloud service or a data base, the buyer can directly be given access from wherever it is. For many who don't have a data storage and need a central server to store the data, this service will also be provided.

2. Automatic metadata generation service: As mentioned in section 3.2, metadata is important for data trading as it serves as a first overview for the dataset offer. Our automatic metadata generation service currently works in two ways, one on the client side and one on the server side. (Fig. 5)

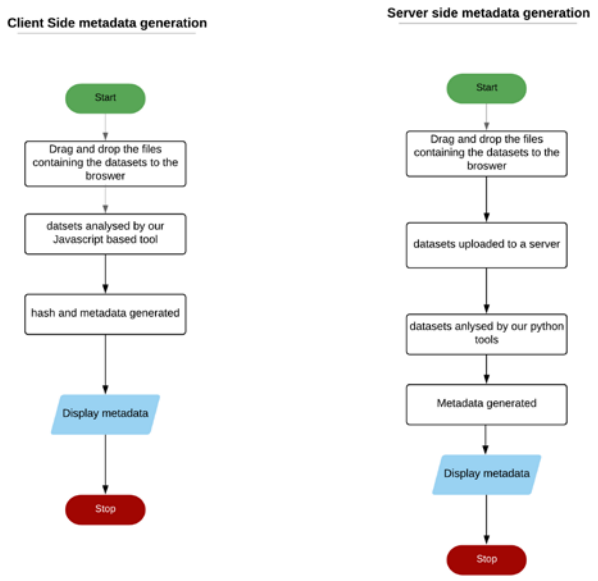

Figure 5. Automatic metadata generation

In the client-side metadata generation method the seller adds the datasets to the browser using the metadata generation service. After the datasets are added in the browser, they are analyzed by our JavaScript based tool. The tool then generates metadata and a hash about the datasets such as number of rows, columns, total number of datasets etc. The tool works only in the browser and no data is duplicated by the system. Although our current tool can generate metadata very precisely, one drawback of this method is that, the trust still lies at the client side. The buyer can still manipulate the metadata before adding it to the data marketplace. Thus, the trust here is on the client-side.

In the server-side metadata generation method, the seller uploads the data to a server. Here the datasets are analyzed by our python tool. Unlike the JavaScript tool, the python tool is more powerful and can analyze metadata more precisely and to a higher extend. But the biggest flaw in this method is that that buyer has to trust the server, that no duplication of data is done.

Both the methods currently have some flaws and specially the server-side method, because it only works when the user uploads the data on our server currently. But this is a base and testing for

our automatic metadata service. Our future work will include solutions for this challenge.

But this way of metadata generation still ensures metadata integrity up to some level and can avoid frauds. After the metadata is generated the selling offer is added to the data marketplace.

The buyer on the other side can look through all the available datasets offer and can select a relevant dataset.

3. Data quality check service: Once the buyer selects the dataset, its added to the shopping cart, now with the extended services the buyer has two options:

- Option 1: check the data quality before buying
- Option 2: Buy data without quality check

If the buyer selects to buy data without checking data quality then, payment service is called.

If the buyer decides to check the quality, the data quality check service is called. At the moment the buyer can choose some criteria from a predefined list, such as extreme values, median and number of *NULL* values. So, the buyer can define how many percent of values out of the range from the extreme values for example (upper and lower minimum) or a how large of variation from the median he is ready to accept. The *Data quality check service* translates these requirements into a query language and checks if this the requirements are fulfilled or not. It is important that the seller cannot see the requirements and the buyer cannot see the data set. For that the algorithm is running in a secure area inside the marketplace. In the next steps we are going to generalize this method in a way that the buyer can describe his requirements on his/her own more powerful language to ensure that he can map all his requirements.

After the quality of the datasets is presented to the buyer. He/she can either proceed with buying the data or can cancel the selection and look for other data sets.

4. Multiple payment options: Just like multiple storage options the system offers multiple payment option such as paying with cryptocurrencies and other financial services. Although cryptocurrencies have many advantages and is being widely adapted now. We realize that some users might prefer conventional banking. Thus, the payment microservices include various payments other than cryptocurrencies such as PayPal, and traditional banking. If the payment is valid, the payment details enter to a smart contract. The smart contract holds the access for the selected dataset and the payment. After all the validation the buyer gets the access to the dataset and the seller gets the expected amount for the datasets. The transaction is then added to the blockchain

5. OPEN CHALLENGES

Although we proposed some methods to achieve some requirements, there are many open challenges with our proposed design and a huge possibility of improvement.

1. Open platform: Although the biggest goal of our approach in to establish a fair, independent data marketplace. We ensured fairness, transparency and removal of middleman using blockchain and smart contracts. The proposed approach is not fully decentralized but rather semi decentralized. The platform and some services still are not decentralized. But ensuring all the functional and non-functional requirements described in the paper and fully decentralized operation of such a platform is still and open challenge.

2. Global data quality: Data quality is a very important characteristic with regards to data and specially data trading. The approach proposed in this paper ensures data quality until some extent, but it still cannot find data quality for any kind of dataset. Ensuring data quality to any dataset is one of the biggest open challenge.

3. Global metadata generation: Both the approaches proposed in this paper can generate precise metadata, but these approaches are not decentralized. Also, the service is not powerful and generic. Finding solutions for generating metadata for any kind of data sets, ensuring integrity of the metadata and the systems having no control on the data is a challenge.

6. CONCLUSION AND FUTURE WORK

Motivated by the massive growth of data usage in recent years and the benefits of blockchain technology, in this paper we present an approach to combine data marketplaces and blockchain for fair and independent data ecosystem. Referring to the research question introduced in section 1, we can state that we made some big Steps forward towards an independent and trustworthy data economy. Our current state of our still ongoing data marketplace is a semi decentralized platform for now. Further our running prototype proves our overall concept and helped us to figure out some more specific challenges.

We also outline some specific requirements such as metadata and data quality associated with data trading and the importance of these properties. In this paper we proposed a semi-decentralized data marketplace based on blockchain and smart contracts. Although the marketplace is semi-decentralized, the control of the data remains in the users hand the system can be trusted. In order to have loosely coupled services our proposed architecture is based on microservices.

Our approach consists of basic services that are required for a data marketplace to function and some extended services which would make data trading more advanced. These services are operated centrally but our current work focuses on finding solutions to make the architecture and the operation of services too as open as possible. But the trade of data is peer-to-peer, the parties just use the marketplace for access transfer and payment. The system has no control over the datasets. The final transaction where the seller gets the expected amount of payment on the buyer get the datasets is carried out by a smart contract. The smart contracts remove the need of a middleman and validates all the transaction and ensures fair exchange.

But in order to fully fair and independent marketplace there are various challenges faced, we identify some major open challenges and present them in this paper. This project is still under process, our future work will include more solutions for still open challenges towards fair data trading.

7. ACKNOWLEDGMENTS

This paper evolved of the research project “Recycling 4.0” (digitalization as the key to the Advanced Circular Economy using the example of innovative vehicle systems) which is funded by the European Regional Development Fund (EFRE | ZW 6-85017297) and managed by the Project Management Agency Bank.

8. REFERENCES

[1] Hasty, A 2015.“Treating Consumer Data Like Oil: How Re-framing Digital Interactions Might Bolster the Federal Trade Commission’s New Privacy Framework,” *Fed. Commun. Law J.*, no. April, pp. 293–324.

[2] Stahl, F., Schomm F., Vossen, G., and Vomfell, L. 2016. “A classification framework for data marketplaces,” *Vietnam J. Comput. Sci.*, vol. 3, no. 3, pp. 137–143.

[3] Zheng, Z., Xie, S., Dai, H., Chen, X., and Wang, H. 2017.“An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends,” *Proc. - 2017 IEEE 6th Int. Congr. Big Data, BigData Congr. 2017*, no. October, pp. 557–564.

[4] Swan, M. 2015. "Blockchain: Blueprint for a new economy." O'Reilly Media, Inc."

[5] Lawrenz, S., Sharma, P. and Rausch, A. 2019. “Blockchain Technology As an Approach for Data Marketplaces,” in P*roceedings of the 2019 International Conference on Blockchain Technology, 2019*, pp. 55–59.

[6] Streamr, “Unstoppable Data for Unstoppable Apps : DATAcoin by Streamr,” *Whitepaper* 2017.[Online available]: https://s3.amazonaws.com/streamr-public/streamr-datacoin-whitepaper-2017-07-25-v1_1.pdf [Accessed: 17-Jan-2020]

[7] Haenni, R. 2020. “Datum White Paper,” *Whitepaper* 2017. [Online available]: https://datum.org/assets/Datum-WhitePaper.pdf [Accessed: 17-Jan-2020]

[8] Lawrenz, S., Sharma, P., and Rausch, A. 2019. “The Significant Role of Metadata for Data Marketplaces,” *Int. Conf. Dublin Core Metadata Appl.*, pp. 95–101.

[9] Newman, S.2019. "Evolutionary Patterns to Transform Your Monolith." O'Reilly Media.