Data Centric DAO: When blockchain reigns over the Cloud

Ibrahim MEHDI*, Moussaab SBAI†, Mohamed MAZLIN‡ and Kamal AZGHIOU§
Data Science & Cloud Computing branch,
École Nationale des Sciences Appliquées Oujda (ENSAO),
UMP - Université Mohammed Premier Oujda (UMP)
Email: *ibrahim.mehdi@ump.ac.ma, †moussaab.sbai@ump.ac.ma, ‡mohamed.mazlin@ump.ac.ma, §k.azghiou@ump.ac.ma

***Abstract*—Nowadays, data has become more and more important. Most governments have passed laws to control how they are owned and used, such as the GDPR in Europe. However, any law is never perfect. We can see it when a company scattered around the world is hampered by the fact of not being able to exploit customer data outside the borders of a territory. Also, we can add that a data owner is not anymore once he sells them to a third party to be exploited. In this work, we propose a solution based on a permissioned blockchain, namely the Hyperledger Fabric, to allow any stakeholder with data to take part, with value providers, in creating Decentralized Autonomous organizations (DAOs). Once created, a DAO can attract other investors to expand it. The Hyperledger Fabric handles this whole process through channels. The infrastructure needed to run the business of the DAO is generated automatically through special transactions taking place at the blockchain level. In the opposite direction, the cloud infrastructure sends notifications to the Hyperledger for traceability and monitoring purposes. Finally, through simulations of wealth distribution models, we show that to keep control of a DAO based on the proposed architecture, the shares must be negotiated according to a Pareto law.**

***Index Terms*—Data-Marketplace, Cloud, Blockchain, Hyperledger Fabric, IoT, DAO, Secure multi-party computation, Data.**

I. Introduction

We will remember 21st century technology for its data-driven gold rush. In fact, the amounts of data do not decrease but increase exponentially. It all started when Sir Tim Berners Lee invented hyperlinks (WWW), which gave life to data sharing around the world. Today, more and more people are using the Internet and new services are appearing, thus generating even more many and diversified data. Nowadays, companies can provide personalized content to each individual user.

Currently, IoT infrastructures generate a large part of the data in the world [1]. Large enterprises are competing to provide the next generation of smart devices to provide their customers with a new lifestyle, as well as to provide enterprises with real-time monitoring of their business operations [2]. Huge futuristic projects are being carried out thanks to the huge amount of connected smart sensors. Smart cities are the perfect example of this, as we are not just envisioning but seeing it solving many problems in urban areas [3]. Sectors such as transport, energy, networks of networks up to public administration all benefit from these solutions. Regardless of the IoT-based system, the underlying infrastructure will always generate a dataset to be mined.

But if you get six good things about the data, you have half a dozen challenges around them. Let us cite the enormous storage capacity necessary to keep the data with a view to their exploitation and/or archiving (Big Data paradigm) and the legal aspect of their exploitation which has sounded the alarm of many organizations in order to protect those of privacy which led them to publish laws that dictate their collection, storage and use [4]. However, these laws do not sufficiently respect the economic aspect of the exploitation of data given the absence of mechanisms allowing their use on a large scale while preserving their private nature. Whichever party may regulate the use of data, whatever the nature of the data, it must not impede technological advances by the measures it imposes.

It is in this context that we introduce in this work a solution that gives more rights to the owner of the data. Indeed, deciding how to exploit the data by giving more control to its owner and guaranteeing him the most interesting return on investment possible are the essential objectives of this work. To prevent the use of data without the knowledge of their owner, we offer a solution based on the encryption of the latter using a secure multi-party calculation. The blockchain makes it possible to keep traces, which are very useful in the event of disputes, thanks to the concept of immutable registers. [5]. To preserve the rights of data users, the architecture proposed in this work stipulates an organizational structure according to the paradigm of Decentralized and Autonomous Organizations (DAO). In fact, a DAO is created as soon as a data owner and a value provider agree. New value provider can join an already created DAO by redistributing the benefits and updating its version. The process repeats itself until the DAO stabilizes on a state of equilibrium which may be due, for example, to an unattractive return on investment for new entrants.

We structure this paper as follows: The section II summarizes the technologies used in our system. The section III outlines some of the related work and explains how our approach is unique. We present the proposed architecture and

978-1-6654-8684-2/22/$31.00 ©2022 IEEE

its components in the section IV. Next, we show how the stakeholder in our system interact with each other in the section V. In the section VII, we cover some aspects of our architecture, along with an explanation of the benefit sharing process. Finally, we end the article with some benefits, general uses of our system, and future work that we can and will implement in our system.

II. Background

In this section, we are presenting an overview about various technologies that are the building blocks for our system.

A. Hyperledger Fabric

1) Overview: Hyperledger Fabric V1.0 was released by the Linux Foundation in 2017. It is an open-source permissioned blockchain framework. It came to life to provide a secure, scalable and flexible groundwork for industrial blockchain solutions. Fabric got rid of the mining process and made sure that access to the data in the ledger is only to authorized members by creating subnets in the network called Channels. Peers can join a channel (or multiple channels at the same time) by enrolling through MSP (Membership Service Provider) and therefore it will have its ledger and chaincodes (smart contracts) installed. Consensus is achieved in Hyperledger Fabric after three phases: Endorsement, Ordering and Validation.

- **Endorsement phase:** Endorsing peers will simulate and execute transactions in an isolated environment and then either sign it as endorsed or not. The result is sent back to the transaction initiator.
- **Ordering phase:** Ordering service (also called ordering service nodes) will receive the transaction and the endorsement signatures and determine the order of transactions.
- **Validation phase:** Validating the authenticity and correctness of blocks of transactions

2) Events in Hyperledger Fabric: There are three sorts of events that can be subscribed to in Hyperledger Fabric:

- **Block events:** Events that are set automatically after commiting a block.
- **Transaction events:** Also set automatically after commiting transaction.
- **Contract events:** Events explicitly added to the chaincode and is set with the contract invocation.

By listening for these events, the application can respond without having to initiate a transaction. In our system, we will use the contract events by setting up an event listener and handler that will allow us to utilize data included in these notifications to automatically command and control an infrastructure.

B. Secure Multi-party computation

The idea of outsourcing data processing and computing without handing over the keys to it is the basis of our system. We know that fully homomorphic encryption [7], [8], allows such privacy, but we also know that it is fully unpractical. It basically goes like this:
Alice encrypts her data $Alice_{Data}$ and sends it $E_{Alice}(Alice_{Data})$ to Bob, does his computation $f(E_{Alice}(Alice_{Data})) = R$ and returns the encrypted result. Alice then decrypts $D_{Alice}(R)$ to find out the results.
Multi-party computation [9] on the other hand is a scheme first developed in 1980's that aims to provide techniques allowing entities to collaboratively calculate a function while keeping their inputs private.
Many protocols are being developed to optimise further the performance of such systems. SPDZ [10], a universal multiparty computing protocol that is safe against up to n-1 of the n players being corrupted by an active attacker, is an interesting one that can be implemented to ensure data privacy.

C. Decentralized Autonomous Organization

A decentralized autonomous organization (DAO) is a collection of entities that work together using smart contracts [11]. This implies that all corporate processes, definitions, and restrictions are encoded in the blockchain and cannot be changed. This type of organization was inspired by the decentralized cryptocurrencies by not having any central authority controlling all the flow and risking both single points of failure and privacy issues. As a result, investors may now purchase DAO shares and get tokens that reflect their ownership in the organization and let them to vote on future initiatives. DAOs are established in our system when two parties agree to collaborate. It's worth noting that one or both of them might already be a DAO.

III. Literature Review

Numerous blockchain based data marketplaces have been introduced in the last few years, each with its own vision and architecture. In [12]–[15] they proposed a decentralized buy/sell architecture based on blockchain as the new solution to traditional data markets. It enforces the fair play between the parties by a punishment system for dishonest behaviours. Elimination of central trusted authorities so that owners of data can have control over it. The implementations of immutable ledgers and smart contracts enabled the users to browse and pick with who they want to work. And to ensure data privacy during the transaction so that no one but the authorized parties can see it, various approaches are presented; Selling the decryption key to an encrypted database [12], or by using swarms in Ethereum network as a decentralized storage for example [13].

In [16]–[18] systems are designed for IoT devices by providing a decentralized platform to sell data streams. These infrastructures are characterized by their low-resources, computational power and lack of security and privacy which requires some sort of middle-system that helps compensate for these weaknesses so they can serve multiple buyers at once. [19] took a step forward with what was already done. A mix of different technologies as hyperledger fabric as an immutable ledger, Cosmos [20] for token interoperability and Mainflux as an IoT gateway. The trade is done by exchanging tokens

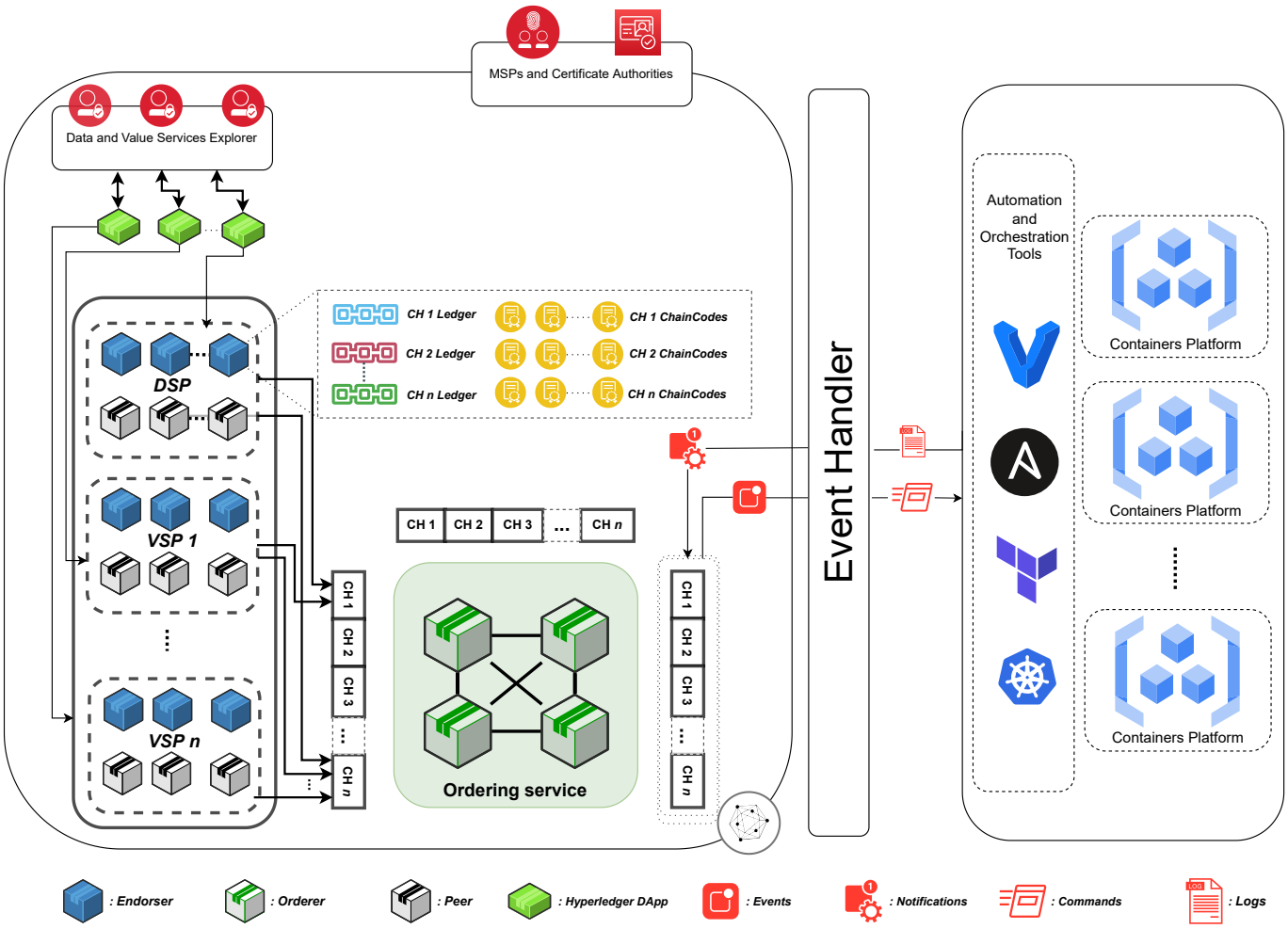

Fig. 1. Data ownership centric architecture

for a Proxied URL with an expiration date determined in the transaction.

We can see in these examples how they all provided a sales marketplace, and at some point the data will be at the other side in plain text with no proof of where it is and how it is being used. Our approach is different. Firstly, we are not interested in selling data as it has already been done before many times. We took the famous sentence ' Data is the new Oil' to the letter. Any person having an oil field can either sell it and be satisfied with the one time payment, which is not what people do, or can be part of the business and invest with his land. As for data, it is important to ensure privacy and anonymity in the exploitation phase. Secondly, we want a decentralized control and monitoring over the infrastructure where all the business is running through blockchain.

IV. Data ownership centric architecture

A. An overview of the proposed architecture

Here, we present an overview of the architecture and its components (Fig. 1). Several parts make up the proposed architecture. Namely, a Hyperledger Fabric module and a Cloud Infrastructure, as well as an event handler facilitating communication between the two by converting Hyperledger events into commands executed by the Cloud and conveying notifications from the latter to the Hyperleder fabric to be registered IV-C. Our goal is to provide a framework for building data-centric DAOs in a non-trust environment. In this way, a data owner would own shares of the DAO concerned as tokens, allowing him to be a member shaping its evolution.

B. The hyperledger part

The Data and Value Services Explorer is where each service is exposed to the various interested parties, with a description containing all the details needed to conclude a business contract. At this level, the stakeholders sign up to benefit subsequently from a space for negotiation in order to launch new DAOs or to join one or more existing ones.

A business contract can be concluded either between one or more data service providers and one or more value providers, or between one or more value providers and an existing DAO. In case of consensus, the stakeholders create a

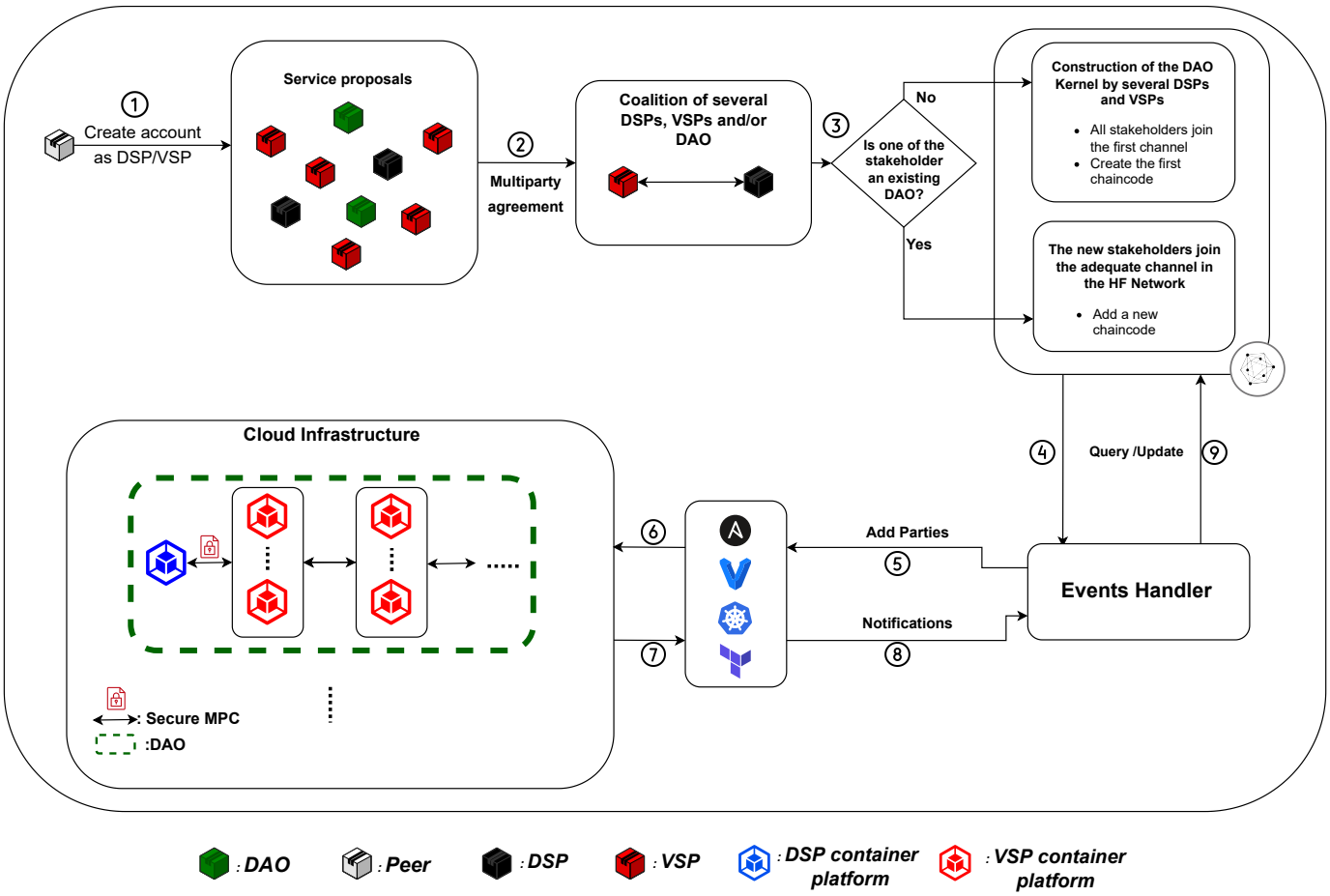

Fig. 2. Interactions between system actors

new DAO or they update the existing DAO.

A concluded contract is materialized by the creation and/or the addition of a new chain in the Hyperledger Fabric where the various transactions will be recorded immutably. Special transactions can auto-generate a cloud infrastructure reflecting the commitment in terms of resources of the stakeholders vis-à-vis the DAO of which they are members. Other types of transactions may concern either the monitoring of the evolution of the DAO or a transfer of rights between an active member and another internal to the DAO or with a new member.

Within each channel, there is a set of blockchain codes that specify the rules of the transactions within it. Then they will send several event notifications to the event handler to command the cloud infrastructure.

A data service provider is the owner of the data, the value provider is any party that can increase the value of the data by subjecting it to a set of operations, such as those relating to machine learning algorithms to derive models for decision making. So the more services above the data layer, the more valuable the DAO.

Channels play a crucial role in maintaining confidentiality within DAOs. In fact, a data service provider DSP and a value provider VSP_1 which form a DAO will be assigned to a channel $CH_1 = \{DSP, VSP_1\}$. This will regulate the workflow using chaincode: profit sharing, approval policies, contract duration and other parameters regarding the business model and the infrastructure ordered. Any other VSP_2 party wishing to join the current DAO will join a new channel CH_2 with the CH_1 still running. Thus, after this last step, the active channels are: $CH_1 = \{DSP, VSP_1\}$, $CH_2 = \{DSP, VSP_1, VSP_2\}$. Each time a new value provider joins an existing DAO, this process will be repeated. In general, the channel CH_i will always have the form of the equation (1):

$$CH_i = \{DSP\} \cup \{VSP_j \ / \ 1 \leq j \leq i\}. \qquad (1)$$

Any party in a given channel will have its ledger and channel codes installed, as well as subsequent channels. For example, the VSP_2 will have the registry and chaincodes from CH_2 up to CH_n. But not CH_1 since he is not a member. That said, we see how the data service provider will be at the heart of the DAO since it takes part in all channels.

C. The infrastructure part

Part of the transactions that take place at the Hyperledger Fabric level concern orders sent to the Cloud infrastructure. In fact, blockchain-level transactions should be able to automatically establish, update, and maintain cloud infrastructure dedicated to running the underlying DAO business processes.

It is by using automation tools [21] such as Vagrant [22], Ansible [23], Terraform [24], Kubernetes [25] that we can launch the instances working in a cloud environment. If a user wishes to work on his own machines, this will be defined in his contract. If no notification comes from his machine, the system assumes he abandoned the job and the contract will be terminated. In sec. VI we will define a future strategy for how to manage and create confidence in the market.

V. Architecture components interactions

A. A chart flow for the proposed architecture

We present in this section, see Fig. 2, the set of interactions between the system components, as well as the flow of actions inside the application.

① A user creates an account so he can be identified, then he selects his user type; Data service provider or a value service provider. Offers are posted on the distributed application with a specific description along with terms and conditions imposed by the original poster. These terms though can be changed as negotiations take place. Users browse and pick their desired partner for a specific job.

② Negotiations are a normal phase before any business agreement. Parties interested in each talk about their terms in order to find a mutual ground, each determines a stake which determines their portion of the DAO.

③ Once everything is set and done, the parties are now considered as one DAO. Now if one of the parties is an existing DAO, the new member joins them in the adequate channel. If not, a new DAO is created combining the two users.

④ Hyperledger fabric will be communicating with the cloud infrastructure using an event handler. It basically reads chaincode events and acts accordingly. This allow us to keep track of all the instances created.

⑤ The event handler should have converted events to script files for our automation tools by this step like Ansible, vagrant, kubernetes and terraform. It will give us the ability to systematise the infrastructure orchestration. ⑥ In this step, the script files are executed and everything is in place. The business will be running inside the cloud infrastructure. ⑦ Frequently, notifications about the big picture of the infrastructure will be sent back into the blockchain so that all members can monitor the flow of work.

⑧ The event handler will be the one converting these notification into queries.

⑨ Queries received are executed and new blocks are created reflecting the current state of the cloud infrastructure.

B. Data privacy

As in [26], a combination of Homomorphic encryption and multiparty computation will allow us to achieve a privacy-preserving framework to make our system even more data centric. It will enable our system to keep the data encrypted all the way through the computations and still have valuable results out of it.

VI. Some aspects of the proposed architecture

As previously stated, our main goal is to make data driven projects preserve the right of privacy with no middle trusted authority. Data will never be decrypted during the whole process. Data owners will be able to make a profit from the final project as our vision is not to sell but to invest with data. One example of the possible business model will be a pay per use or subscription based, it all depends on the contract between the providers.

IoT infrastructures will rely significantly on our technology to offer a wide range of data to other parties for analysis and use while maintaining data privacy.

In terms of profit splitting, our system distributs profit π group wise and according to each party's stake as defined in the chaincodes of each channel. See Fig. 3 for a detailed tree graph describing the ownership percentage in a given DAO. The (i) in $DAO^{(i)}$ describes the version of the DAO created in channel CH_i. In order to compute the profit of a specific $DAO^{(i)}$ or VSP_i, we use the equations 2 and 3.

$$\pi[DAO^{(i)}] = \pi[DAO^{(n)}] \times \prod_{k=i+1}^{n} \beta_k \tag{2}$$

$$\pi[VSP_i] = (1 - \beta_i) \times \pi[DAO^{(i)}] \tag{3}$$

For stakeholders to build or integrating a DAO, the various stakeholders must agree on ratios β_i for the distribution of wealth. This value is used to calculate a member's share of total profit.

To visualize the evolution of the shares of the data owner as well as those of the value providers integrating the DAO step by step, we present in Fig. 4 and Fig. 5 graphs for two different scenarios: (i) In the first scenario we let's simulate the β_i from a uniform law, (ii) In the second scenario the β_i are taken from the equation 3 on which we have imposed a distribution of wealth according to the law of Pareto. We inject the deducted β_i into the equation 2 to generate the profits, as shown in the Fig. 5.

1) Randomly selected values: Creating β_i with this method is straightforward. Generate uniformly distributed and sorted values in an interval. We started with $25 - 99\%$ then we increased it gradually and noticed that the value of all β_i should be at least 70% (in our example of 7 $VSPs$) in order to keep the Data Owner at the top of the pile(see Fig. 4). We also notice how the VSPs will have more leverage if they join a job at the end of it. This will make VSPs refuse to join works

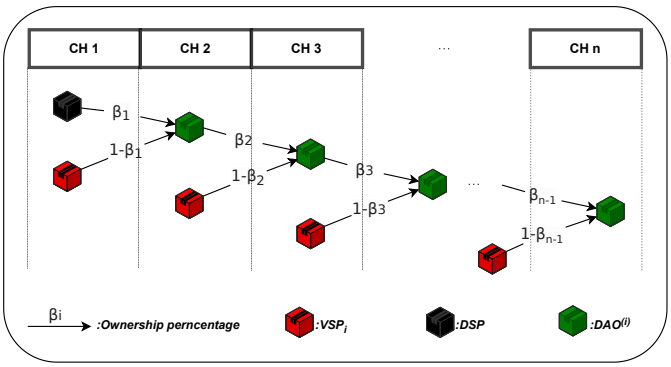

Fig. 3. Ownership percentages of each member in corresponding DAO

in its beginning. It can be solved by changing the range of β_i $80-99\%$ but it will bring more problems than it solves as the DSP would monopolise the whole business. In this approach, we created β_i then analyzed the total ownership percentage.

2) Selecting values according to Pareto Principle: The Pareto principle [27], named after Vilfredo Pareto, an Italian economist and sociologist. It was developed to describe the distribution of majority of wealth in the hands of a top percentage of the population. It has many uses in multiple areas besides economy as insurance, manufacturing, management and many more. We will be using it to generate numbers following the distribution in question representing the total ownership percentage.

Algorithm Generate β_i fitting the total ownership to a Pareto distribution

Require: $n \geq 2$ ▷ Number of β_i

1: Generating n random numbers following Pareto's distribution.
2: Scale (between 0 and 1 to describe percentage) then sort descending these numbers.
3: Assign each value with corresponding member starting with DSP and the highest value. ▷
Now we have for each member his total percentage. We can calculate β_i by reversing formulas 3

As we can see in Fig. 5, the percentage of total ownership drops down as we add more VSPs. This is a better result than what we got before. It is logical that as long as we add more VSPs, the value of the DAO will increase which leads to higher share prices. Contrary to what we did on the first one, in this approach we defined the total ownership percentage then deduced the β_i.

VII. Conclusion

Our system presents a functioning framework that allows anyone, and especially IoT infrastructure owners to invest their data in a DAO amongst other service providers to further increase the value extracted from that data. This will benefit IoT device manufacturers since our future consists of more gadgets and sensors all over the cities and infrastructures, analytics services by having broader data point thus reducing the effects of 'small dataset curse' also known as overfitting, having real-time data to stay up to date. And the end users of course that will use the final service. Our motivation is to have always our data protected but still making it work in the real world to improve services, applications and overall user experience. As a future development, a user reputation system may be implemented. Parties can offer feedback on their interactions with a specific user by including a quality of experience mechanism, which will help us sanction poor conduct. This will boost user confidence, which is an important component of the online experience.

References

[1] Keyur K Patel, Sunil M Patel, et al. Internet of things-iot: definition, characteristics, architecture, enabling technologies, application & future challenges. *International journal of engineering science and computing*, 6(5), 2016.

[2] Manlio Del Giudice. Discovering the internet of things (iot) within the business process management: A literature review on technological revitalization. *Business Process Management Journal*, 2016.

[3] Veronica Scuotto, Alberto Ferraris, and Stefano Bresciani. Internet of things: applications and challenges in smart cities. a case study of ibm smart city projects. *Business Process Management Journal*, 2016.

[4] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[5] Massimo Di Pierro. What is the blockchain? *Computing in Science & Engineering*, 19(5):92–95, 2017.

[6] Elli Androulaki, Artem Barger, Vita Bortnikov, Christian Cachin, Konstantinos Christidis, Angelo De Caro, David Enyeart, Christopher Ferris, Gennady Laventman, Yacov Manevich, et al. Hyperledger fabric: a distributed operating system for permissioned blockchains. In *Proceedings of the thirteenth EuroSys conference*, pages 1–15, 2018.

[7] Joon-Woo Lee, HyungChul Kang, Yongwoo Lee, Woosuk Choi, Jieun Eom, Maxim Deryabin, Eunsang Lee, Junghyun Lee, Donghoon Yoo, Young-Sik Kim, et al. Privacy-preserving machine learning with fully homomorphic encryption for deep neural network. *IEEE Access*, 10:30039–30054, 2022.

[8] Craig Gentry. *A fully homomorphic encryption scheme*. Stanford university, 2009.

[9] Oded Goldreich. Secure multi-party computation. *Manuscript. Preliminary version*, 78:110, 1998.

[10] Ivan Damgård, Marcel Keller, Enrique Larraia, Valerio Pastro, Peter Scholl, and Nigel P Smart. Practical covertly secure mpc for dishonest majority–or: breaking the spdz limits. In *European Symposium on Research in Computer Security*, pages 1–18. Springer, 2013.

[11] Alexandra Sims. Blockchain and decentralised autonomous organisations (daos): The evolution of companies? 2019.

[12] Matias Travizano, Carlos Sarraute, Gustavo Ajzenman, and Martin Minnoni. Wibson: A decentralized data marketplace. *CoRR*, abs/1812.09966, 2018.

[13] Kazim Rifat Özyilmaz, Mehmet Doğan, and Arda Yurdakul. Idmob: Iot data marketplace on blockchain. In *2018 crypto valley conference on blockchain technology (CVCBT)*, pages 11–19. IEEE, 2018.

[14] Hyunkyung Yoo and Namseok Ko. Blockchain based data marketplace system. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1255–1257. IEEE, 2020.

[15] Prabal Banerjee and Sushmita Ruj. Blockchain enabled data marketplace - design and challenges. *CoRR*, abs/1811.11462, 2018.

[16] Pooja Gupta, Volkan Dedeoglu, Salil Kanhere, and Raja Jurdak. Towards a blockchain powered iot data marketplace. pages 366–368, 01 2021.

[17] Ahmed Suliman, Zainab Husain, Menatallah Abououf, Mansoor Alblooshi, and Khaled Salah. Monetization of iot data using smart contracts. *IET Networks*, 8(1):32–37, 2019.

[18] Krešimir Mišura and Mario Žagar. Data marketplace for internet of things. In *2016 International Conference on Smart Systems and Technologies (SST)*, pages 255–260, 2016.

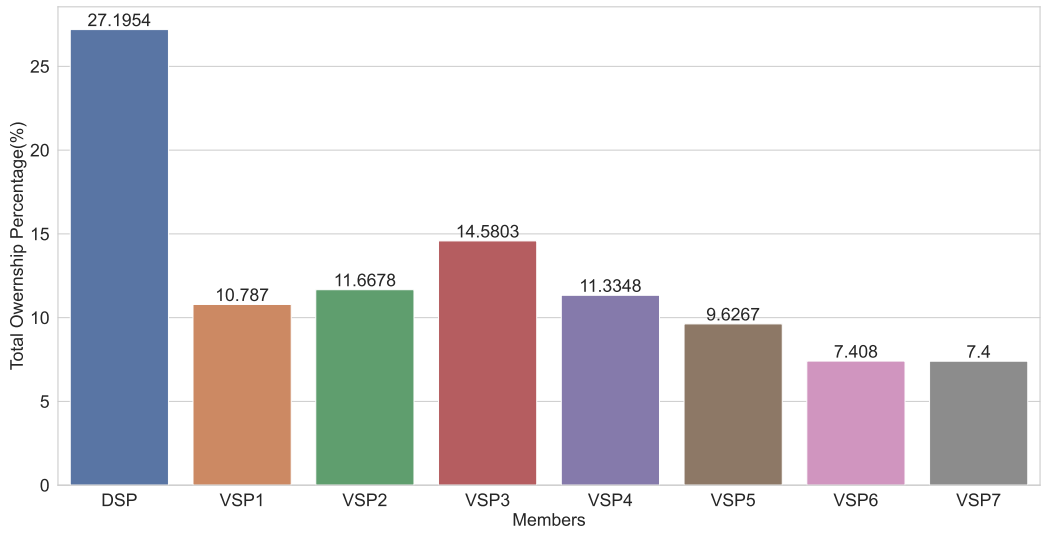

Fig. 4. Total Ownership percentage calculated with uniformly random β_i values using formula 3. Results may vary with randomness in play, but doesn't change the inefficiency of such distribution method.

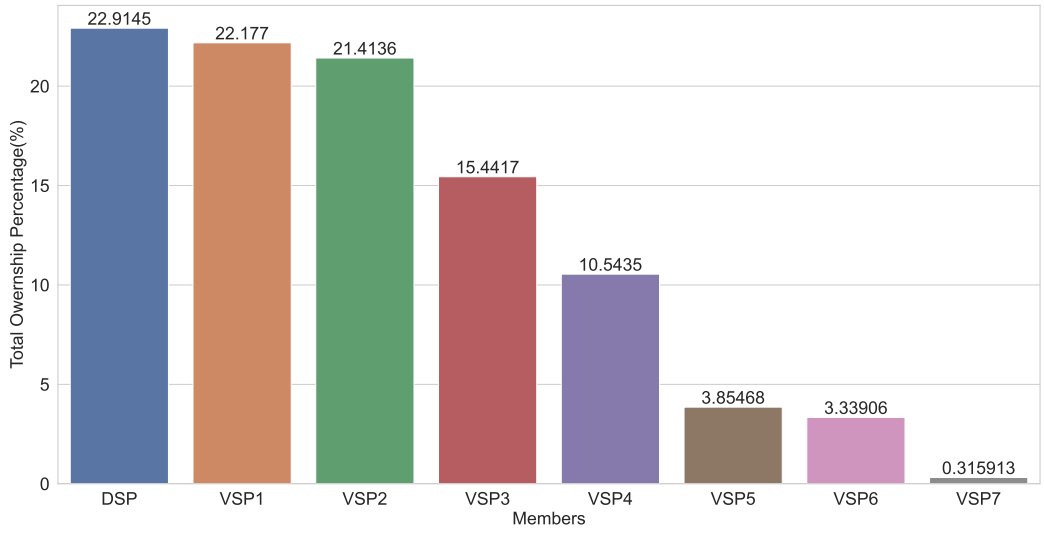

Fig. 5. Total Ownership following the Pareto distribution approach. A remarkable consistency and fairness using β_i fitting to that distribution.

[19] Drasco Draskovic and George Saleh. Datapace-decentralized data marketplace based on blockchain. *Datapace*, 2017.

[20] Jae Kwon and Ethan Buchman. Cosmos whitepaper. *A Netw. Distrib. Ledgers*, 2019.

[21] Loic Houde, Daniel Jacob, Tovo Rabemanantsoa, and Jean-François Rey. *Gestion Automatique d'Environnement Virtuel (GAEV)*. PhD thesis, INRAE, 2021.

[22] Mitchell Hashimoto. *Vagrant: up and running: create and manage virtualized development environments*. " O'Reilly Media, Inc.", 2013.

[23] Lorin Hochstein and Rene Moser. *Ansible: Up and Running: Automating configuration management and deployment the easy way*. " O'Reilly Media, Inc.", 2017.

[24] Yevgeniy Brikman. *Terraform: up & running: writing infrastructure as code*. O'Reilly Media, 2019.

[25] Brendan Burns, Joe Beda, and Kelsey Hightower. *Kubernetes: up and running: dive into the future of infrastructure*. O'Reilly Media, 2019.

[26] Ivan Damgård, Valerio Pastro, Nigel Smart, and Sarah Zakarias. Multiparty computation from somewhat homomorphic encryption. In *Annual Cryptology Conference*, pages 643–662. Springer, 2012.

[27] Rosie Dunford, Quanrong Su, and Ekraj Tamang. The pareto principle. 2014.