DLMetaChain: An IoT Data Lake Architecture Based on the Blockchain

Michalis Pingos
Department of Electrical Engineering,
Computer Engineering and Informatics
Cyprus University of Technology
Limassol
michalis.pingos@cut.ac.cy

Panayiotis Christodoulou
Department of Computer Science

Neapolis University
Pafos
panayiotis.christodoulou@nup.ac.cy

Andreas Andreou
Department of Electrical Engineering,
Computer Engineering and Informatics
Cyprus University of Technology
Limassol
andreas.andreou@cut.ac.cy

Abstract— **Nowadays, the IoT ecosystem is evolving rapidly, with multiple heterogeneous sources producing high volumes of data and processes transforming this data into meaningful or "smart" information . These volumes of data, including IoT data, need to be stored in repositories that can host raw, unprocessed, relational and non-relational types of data, such as Data Lakes. Due to the weakness of metadata management, security & access control is one of the main challenges of Big Data storage architectures as Data Lakes can be replaced without oversight of the contents. Recently, the Blockchain technology has been introduced as an effective solution to build trust between different entities, where trust is either nonexistent or unproven, and to address security and privacy concerns. In this paper we introduce DLMetaChain, an extended Data Lake metadata mechanism that consists of data from heterogeneous data sources which interact with IoT data. The extended mechanism mainly focuses on developing an architecture to ensure that the data in the Data Lake is not modified or altered by taking into advantage the capabilities of the Blockchain.**

Keywords— Internet of Things, Smart Data Processing, Data Lakes, Heterogeneous Data Sources, Metadata Mechanism, Data Blueprint, Blockchain, Smart Contracts

I. INTRODUCTION

Big Data has been called "the new oil" as it is recognized as a valuable human asset, which, with the proper collation and analysis can deliver information that will give birth to deep insights into many aspects of our everyday life and, moreover, to let us predict what might happen in the future. Big data is essentially a combination of structured, semi-structured and unstructured data that originate mostly from one of five primary sources: media, Cloud, Web, traditional business systems and Internet of Things (IoT) [1].

The amount of data produced and communicated over the Internet and the Web is rapidly increasing [2]. Every day, around 2,5 quintillion bytes of data are produced. There were approximately 44 zettabytes of data in the world in 2020. Given how much data is created every day, there will likely be 175 zettabytes and 75 billion Internet-of-Things (IoT) devices in the world by the year 2025 [3]. This data include textual content such as unstructured, semi-structured, and structured data to multimedia content such as images, video, and audio on a variety of platforms such as enterprise, social media, and sensors [2].

Despite the great and drastic solutions proposed in recent years in the area of Big Data Processing and Systems of Deep Insight, the treatment of Big Data produced by multiple heterogeneous data sources remains a challenging and unsolved problem. A Data Lake (DL) is a repository that can store a large amount of structured, semi-structured, and unstructured data. It is a place to store every type of data in its native format with no fixed limits on account size or file, and offers high data quantity to increase analytic performance and native integration. A DL is a quite new data storage architecture linked with Big Data processing with unsolved challenging problems [4]. Two of the major and challenging problems of DL are the following: (i) there is no descriptive metadata or mechanism to maintain metadata leading to data swamp [5], and, (ii) security (privacy and regulatory requirements) and access control as data in a lake can be replaced without the oversight of the contents.

Blockchain's purpose is to offer secure and transparent ways to record and transfer data [6] [22]. In this paper we propose a framework, namely DLMetaChain, that uses the Blockchain technology to protect sensitive and crucial IoT data and to ensure that the data in the DL is not modified. The aim of this paper is to create a DL structure (ponds or zones) that cannot be altered so as to prevent fraud and unauthorized activity between DL of different organizations. The Blockchain technology also enhances privacy issues by anonymizing personal data and also is used to grant authorized access to DL owners.

The remainder of the paper is structured as follows: Section II discusses related work and the technical background in the areas of DL, IoT and Blockchain. Section III presents the DLMetaChain IV framework along with its main components, while, in section authors evaluate the proposed architecture and present the experimental results . Finally, section V concludes the paper.

II. RELATED WORK/TECHNICAL BACKGROUND

The area of smart data processing comprises the ability to clearly define, interoperate, openly share, access, transform, link, syndicate, and manage data. Under this perspective, it becomes crucial to have various knowledge-based metadata representation techniques to structure data sets, annotate them,

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

link them with associated processes and software services, and deliver or syndicate information to recipients. The Smart Data Processing Systems area can include various topics to fully utilize the aforementioned capabilities, such as data ingestion, data aggregation of an enormous variety of structured, unstructured and semi-structured datasets, knowledge-based meta-data representation techniques for the conversion of raw into smart data, AutoML process techniques, data privacy and protection, automated deployment, run-time software performance monitoring and dynamic configuration [1] [7].

In addition, this area includes adaptive frameworks and tool-suites that support smart data processing by using both data in motion (e.g. data streams from sensors), and data at rest, that rely on advanced techniques for efficient resource management, and partitioning of intensive data workloads across a number of private and public clouds. Smart data processing supports the process and integration of data into a unified view from disparate Big Data sources including Hadoop and NoSQL, DL, data warehouses, sensors and devices in the Internet of Things, social platforms, and databases, whether on-premises or cloud, structured or unstructured and software-as-a-service applications to support Big Data analytics [7].

Common fields of data processing systems are semantic models, structured data configurations, DL, data warehouses, Machine Learning (ML) and ontologies. One of the most significant findings in these studies are the importance of using the DL architecture to store large amounts of relational and non-relational data combining them with traditional data warehouses. Another notable finding is the exploitation of ontology frameworks in order to manage and make heterogenous data sources that produce large amounts of data, meaningful. Finally, another major finding of that work is the need for exploitation of Machine Learning (ML), and especially AutoML, which focuses on automating repetitive tasks of ML. Various papers incorporate DL, knowledge-based meta-data ontologies, ML and AutoML to tackle data processing issues.

Most of the work conducted on Big Data integration has been focused on the problem of processing very large sources, extracting information from multiple, possibly conflicting data sources, reconciling the values and providing unified access to data residing in multiple, autonomous data sources. Various studies mainly addressed isolated aspects of data source management relying on schema mapping and semantic integration of different sources [8] [9]. Those studies focused mostly on the construction of a global schema or a knowledge base to describe the domain of the data sources. Web table search is also closely related to data source search. Most of the proposed techniques outlined examine user queries and return tables related to specific keywords presented in the query [8] [10] [11]. However, keyword-based techniques fail to capture the semantics of natural language, i.e., the intentions of the users, and thus they can only go as far as giving relevant hits.

DL is one of the arguable concepts that appeared in the era of Big Data. The idea of a DL originates from the business field instead of academic. As reported in [12], a DL is a quite new data storage architecture linked with Big Data processing with unsolved challenging problems such as:

- DL cannot determine data quality or the lineage of findings
- DL accept any data without oversight and governance
- There is no descriptive metadata or mechanism to maintain metadata leading to data swamp
- Data need to analyse from scratch every time
- Performance cannot be guaranteed
- Security (privacy and regulatory requirements) and access control (weakness of metadata management) as data in a lake can be replaced without oversight of the contents

The research and business communities showed great interest and carried out satisfactory research work on DL. Despite this fact, many of the issues require considerable effort to achieve the desired level of DL utilization in the area of Big Data and Business Intelligence. A challenging open research issue is also the lack of an existing DL framework that provides standardization policy and a metadata mechanism that can treat effectively and efficiently Big Data, including IoT data coming from different heterogeneous data sources producing different types of data before the ingestion in a DL and before the extraction of the knowledge and information from the DL, providing also security and privacy of the data.

The authors in [5], identified and presented six main functional characteristics that should ideally be provided by a DL metadata system:

- Semantic Enrichment (SE)
- Data Indexing (DI)
- Link generation and conservation (LG)
- Data Polymorphism (DP)
- Data Versioning (DV)
- Usage Tracking (UT)

The work in [13] extended the aforementioned list of characteristics by comparing the metadata mechanism with the two most completed systems as presented in [14]: CoreKG [15] and MEDAL [14]. The new list of the characteristics include:

- Granularity
- Ease of storing/retrieval
- Size and type of metadata
- Expandability

None of the existing papers mentions privacy and security as a characteristic that can add value to the synthetic examination of the quality and efficiency of metadata enrichment mechanisms for DL.

On the other hand, nowadays the Blockchain technology becomes more and more popular and are gradually becoming part of the infrastructure and are paving the way for novel applications [21]. Blockchain provides a distributed p2p communication network where non-trusting nodes can interact with each other without a trusted intermediary. In a more verifiable manner, Blockchain is a decentralized database located on a p2p network that has its own protocols and offers traceability, transparency and privacy and security.

The goal of the work reported in [16] is to introduce a Blockchain-based access-control manager to health records to address the industry interoperability challenges expressed in the Office of the National Coordinator for Health Information Technology's (ONC) and shares a Nationwide Interoperability Roadmap in which all data must be stored in a DL.

The authors in [17] present the benefits and challenges of integrating Blockchain with IoT. According to this paper, this combination brings many advantages, such as publicity, decentralization, resiliency, security, speed, cost saving and immutability, which improve many of the IoT issues. At the same time, it introduces new challenges that should be addressed, such as scalability, process power and time, storage, lack of skills, legal and compliance and naming and discovery.

A decentralized application called ParkChain is described in [18], which is based on two emerging technologies: IoT and Blockchain. The Blockchain applied to the ParkChain system prevents an unauthenticated user from entering a controlled parking space. The preliminary results of ParkChain verify that Blockchain enables trusted access with low cost in terms of cost and time.

According to [19], in order for the manufacturing industry to be modernized, it should adopt digital twin technology within their operations, outputs, and offerings. Since then, the digital twin paradigm's distinct contributions have increased significantly as a result of seamless synchronization with a number of cutting-edge technologies such as the Internet of Things (IoT), artificial intelligence (AI), big and streaming data analytics, DL, software-defined cloud environments, Blockchain, and so on.

DL data storage is essential when handling IoT data produced by multiple heterogeneous data sources, such as Social Media, Cloud, Web and Business Systems as demonstrated also in most of the papers mentioned in this section (see figure 1).

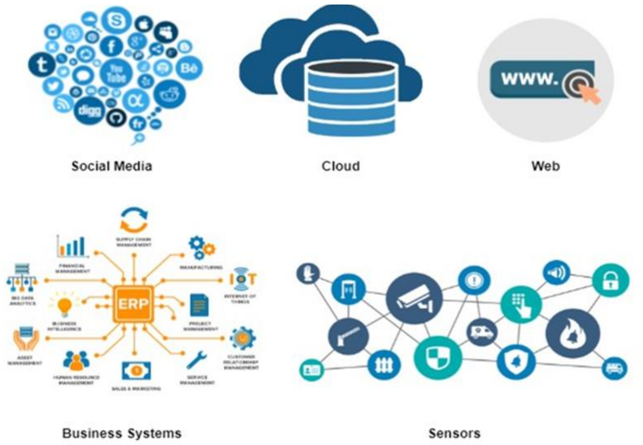


Figure 1 The five primary types of data sources

This paper extends and enhances previous work on the topic [13] which adopts the basic principles of manufacturing blueprints [20] and modifies their purpose and meaning to reflect the description and characterization of sources and the data they produce via the utilization of the five Big Data characteristics (5Vs - see figure 2). These characteristics essentially describe data sources by means of specific types of blueprints through an ontology-based description representation. Big Data sources will thus be accompanied by a blueprint metadata description before they become part of a DL. The latter follows a pond architecture.

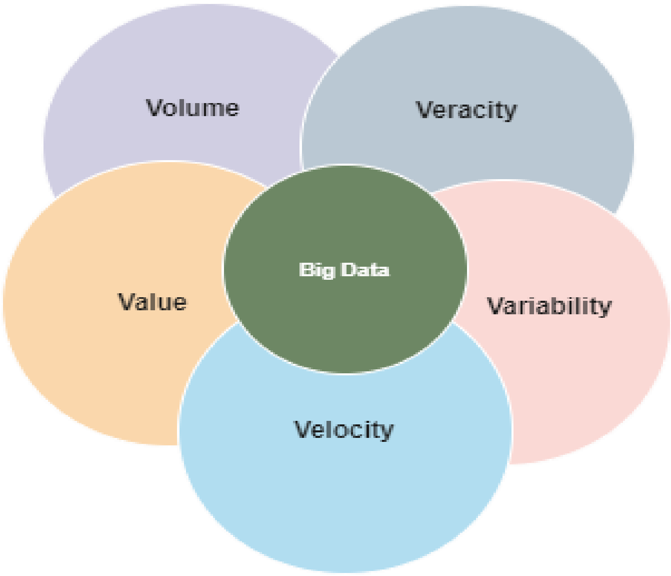


Figure 2 The 5V's of big data

III. DLMETACHAIN FRAMEWORK ARCHITECTURE

The goal of the proposed framework is to enrich the existing novel DL pond metadata mechanism framework described in [13] with the Blockchain technology so as to provide more security and privacy and to ensure that the data in the DL have not been modified or altered.

A. DL Blueprint

According to [13] each candidate source, including IoT data sources, needs to be characterized first by a metadata mechanism namely Data Source Blueprint (DSB) before it becomes a member of a DL [13]. This characterization is performed with the contribution of the 5 basic characteristics of Big Data, namely Volume, Variety, Veracity, Value, and Velocity. The blueprint mechanism is divided into two parts, the stable data blueprint, and the dynamic data blueprint. The dynamic blueprint consists of attributes that may change during the process of data processing or while a data source generates new data (see figure 3). Essentially, the dynamic and stable blueprint which form the DSB is an RDF (Resource Description Framework) file following the XML structure. Figure 3 presents the value types of each attribute.

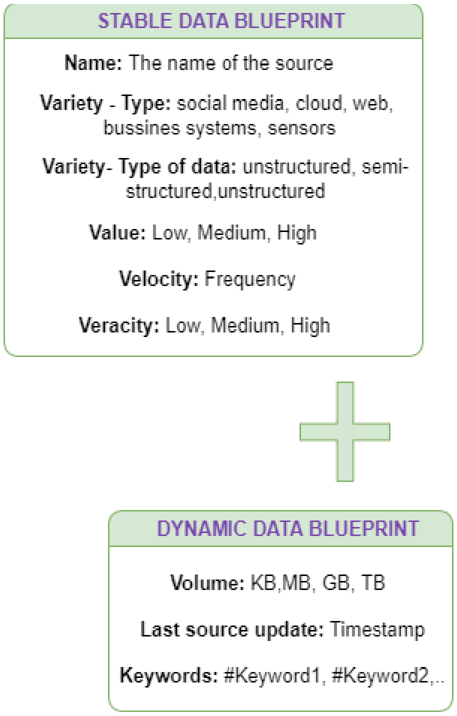

STABLE DATA BLUEPRINT
Name: The name of the source
Variety - Type: social media, cloud, web, bussines systems, sensors
Variety- Type of data: unstructured, semi-structured,unstructured
Value: Low, Medium, High
Velocity: Frequency
Veracity: Low, Medium, High
DYNAMIC DATA BLUEPRINT
Volume: KB,MB, GB, TB
Last source update: Timestamp
Keywords: #Keyword1, #Keyword2,..

Figure 3 DSB attributes of stable and dynamic blueprint

For example, let us assume that we want to select healthcare data sources including IoT data to add them in a DL. This can be done by selecting among the candidate sources which bear the characteristics according to the DSB as presented in figure 3:

- Source 1

  Stable Blueprint Attributes:

  Name: Activity phone sensor

  Variety-Type: Sensor

  Variety-Type of data: semi-structured

  Value: High

  Velocity: 50ms

  Veracity: High

  Dynamic Blueprint Attributes:

  Volume: KB

  Last Update: 01/01/2022; 10:35

  Keywords: # Activity, #Acceleration #HeartHealth

- Source 2

  Stable Blueprint attributes:

  Name: Phone users

  Variety-Type: Business Systems

  Variety-Type of data: structured

  Value: High

  Velocity: 360h

  Veracity: High

  Dynamic Blueprint attributes:

  Volume: KB

  Last Update: 01/01/2022 19:40

  Keywords: #PhoneUsers #Activity #HeartHealth

- Source 3

  Stable Blueprint attributes:

  Name: Blood Cell Images

  Variety-Type: Cloud

  Variety-Type of data: Unstructured

  Value: High

  Velocity: Monthly

  Veracity: High

  Dynamic Blueprint attributes:

  Volume: MB

  Last Update: 24/02/2022 06:50

  Keywords: #HeartHealth #BloodCells

As previously mentioned, the DSB is an RDF file that follows the XML structure. The following RDF representation presents the DSB of Source 1:

Stable Blueprint

```
<?xml version="1.0">
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"
xmlns:si="blueprints.com/">
<rdf:Description
rdf:about="blueprints.com/stable">
<cd:name>Activity phone sensor</cd:name>
 <cd:varytype>Sensor</cd:varietytype>
 <cd:varytypf> semi-structured</cd:varytpf>
 <cd:value>High</cd:value>
 <cd:velocity>50ms</cd:velocity>
 <cd:veracity>High</cd:veracity>
</rdf:Description>
</rdf:RDF>
```

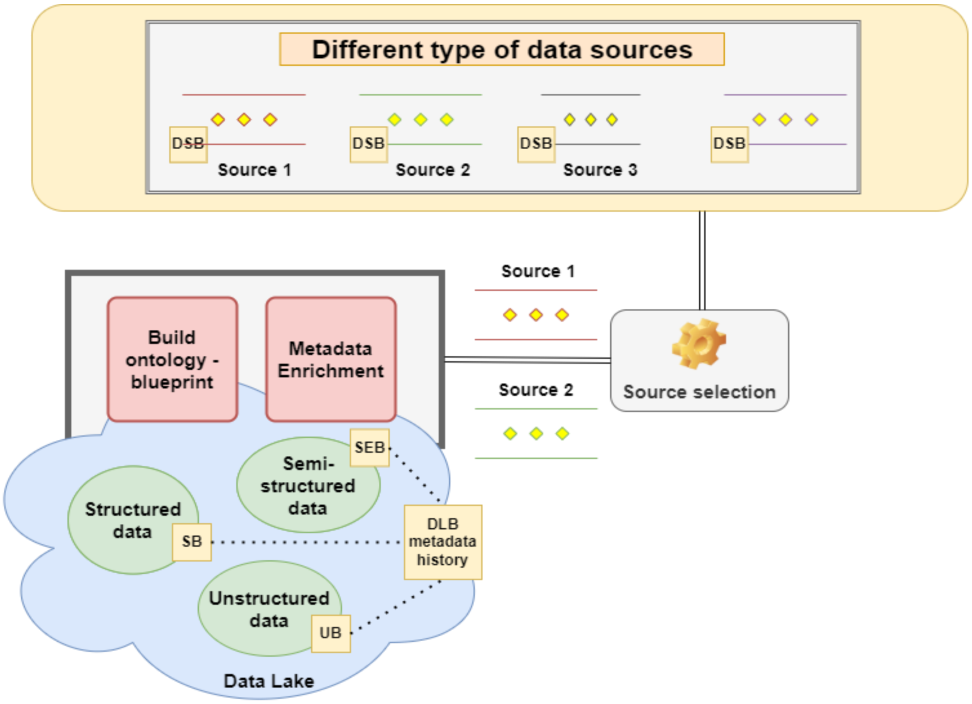


Figure 4 DLMetachain framework architecture DLB metadata history creation

Dynamic Blueprint

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"
xmlns:si="www.blueprints.com">

<rdf:Description
rdf:about="blueprints.com/dynamic">

 <cd:volume>KB</cd:volume>
 <cd:lastupdt>24/02/2022;10:35</cd:lastupdt>
 <cd:keywords>
    <li>Activity</li>
    <li>Acceleration</li>
    <li>Heart Health</li>
 </cd:keywords>
</rdf:Description>
</rdf:RDF>
```

In order to select which data sources will be part of a DL, dedicated middleware (Figure 4) runs a query similar to the one below:

```
SELECT ? sources
        WHERE {
        ? source <has value>   High    &&
                  <has veracity> Medium &&
                   <has keyword> #Activity
                }
```

After execution of the SPARQL query various sources become members of a DL. In our case Source 1 and Source 2 become members of the DL, while Source 3, which does not satisfy the query, do not. Therefore, the stable and dynamic blueprint of the two selected sources incorporate now the DLB metadata history as shown in Figure 4. Source 1 has become a member of the DL pond with unstructured data, while Source 2 of the pond with structured data due to the Variety-Type attribute in the stable blueprint of these data sources.

Finally, both the dynamic and stable blueprints of the selected sources are stored in a SB (structured data blueprint) and a UB (unstructured data blueprint) (see figure 4) which constitute the DLB metadata history. Each time a new source is pushed in the DL, or each time a source generates new data that modify the dynamic blueprint, a new version of the SB and UB is created, thus the DLB metadata history changes (Figure 4).

B. Smart Contract

To ensure that the DLB metadata will not be altered or modified a Blockchain smart contract is developed within the proposed framework that has a two-fold purpose, (i) to allow a DL owner to register a DL into the Blockchain based on its metadata, and, (ii) to allow end-users to verify the correctness of a DL before conducting any actions on it.

As outlined in Figure 5, each time a new version of a DLB metadata is created a new SHA256 value for that specific DLB is generated automatically and is used along with the DL's *id* to register the specific DL into the Blockchain via the proposed smart contract. Multiple versions of a DLB can be created for each DL that results in the creation of a unique DL Chain for each DL and a whole new Chain for the full system, namely, DLMetaChain.

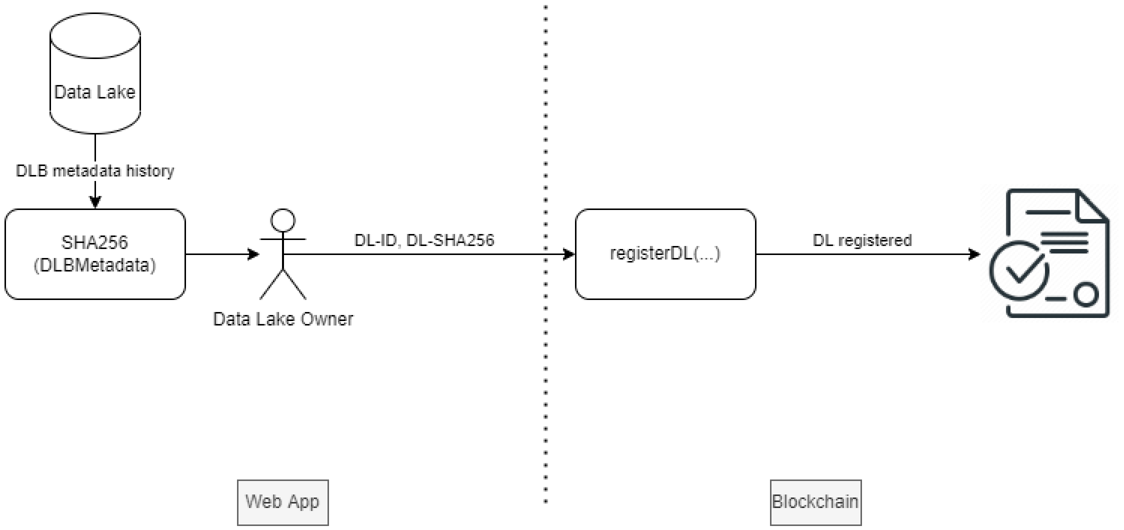


Figure 5 DLMetachain framework architecture DLB metadata history creation

As previously mentioned, the purpose of the proposed smart contract is twofold, thus, mechanisms were developed that can be used by end-users to verify that a shared DL has not been modified or altered.

Figure 6 presents an overview of the system from the end-user's perspective. Firstly, the DL owner shares a DL *id* to an end-user who uses it to retrieve DL metadata. Once the metadata is retrieved, a mechanism that generates the SHA256 value for that specific source is executed and then the user uses that value to check if the DL exists on the Blockchain or not. If the DL was successfully retrieved by the smart contract then this is a proof that the DL has not been modified, thus the user can execute actions using the specific DLB. If the DL cannot be found on the smart contract, then this means that the DL was compromised and the end-user is recommended not make use of the DL.

IV. USE CASE SCENARIO

To demonstrate the effectiveness and usage of the proposed framework we have deployed the recommended smart contract on the Ethereum Rinkeby Test Network. The list of all executed transactions, as well as the source code of the smart contract can be found on the following smart contract address 0x0E864521Ccf8BD65aBFcC920ec43d93fdf82D80a. The ETH public address 0x6c56618BCbF502b237369551cF2Af7317E763eDb, acts as the Administrator of the developed dApp, thus it can register DLs into the smart contract.

Before evaluation, the `registerDL(…)` function of the smart contract firstly, two different versions of a DL were created ,that can be found on our GitHub repository, using IoT data and then the SHA256 value for each one of the versions was generated. The SHA256 values of the DL versions are shown below:

DL0_v1: 5a221f47e54beac4c9548116bf9196a23b041c0c664280d018c96a1089643568

DL0_v2:

db9772aefdaa674d0c4b3ebc61a30d2409a8a07f41925a9b2b4192490ce98530

Now as the SHA256 values of the DLs were generated the Administrator of the proposed framework can register them to the smart contract by calling the `registerDataLake(…)` function and by providing the DL *id* along with its SHA256 value.

To test the efficiency of the proposed Use Case scenario we have registered both versions of the DL into the smart contract. The cost for deploying the smart contract as well as the cost for registering a DL version on the smart contract can be on Table 1.

Table 1 Cost of deploying and registering DL

Deployment/Functions	**Cost (ETH)**	**Cost (USD)**
Contract Deployment	0.001822	6.19
registerDataLake()	0.000121	0.41

Furthermore, as the DL was successfully registered into the smart contract, the Administrator of the dApp can share the source code of the DL with an end-user.

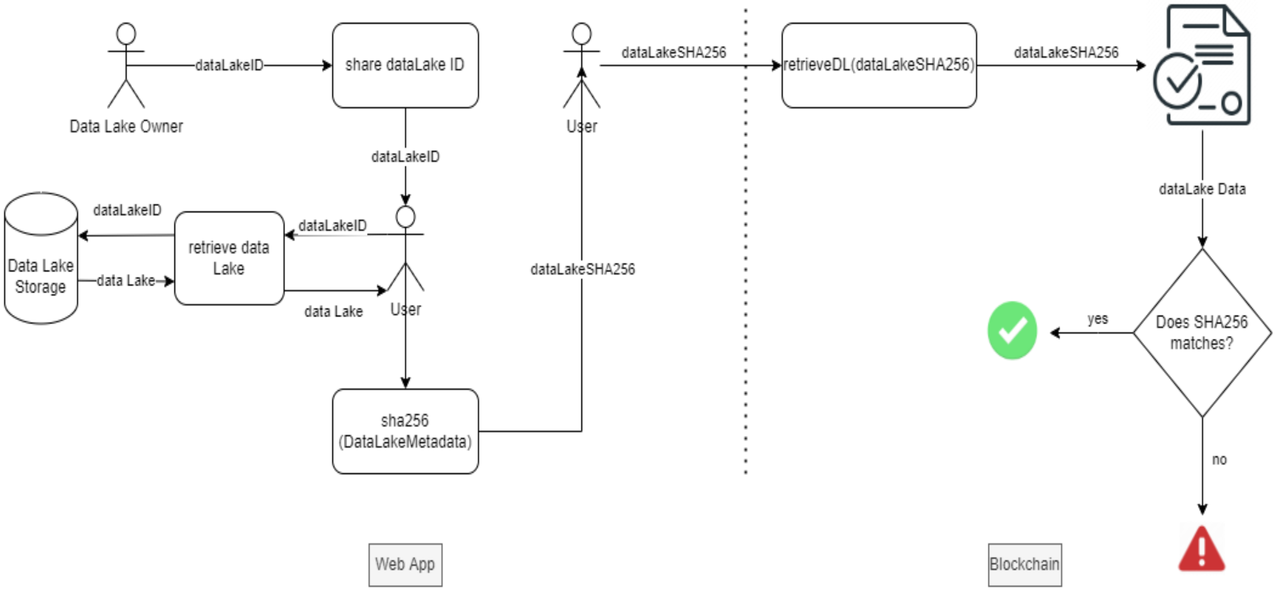
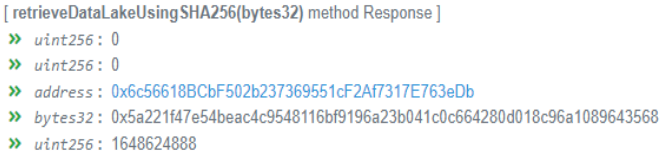


Figure 6 Overview of the system via the end-user's perspective

When the end-user wishes to check that the DL was not compromised, she firstly uses the proposed dApp to find the SHA256 value of the shared DL and then calls the `retrieveDataLakeUsingSHA256(…)` function to check whether the DL exists or not. To demonstrate the effectiveness of the proposed approach, the specific function was called using as input the value 0x5a221f47e54beac4c9548116bf9196a23b041c0c664280d018c96a1089643568 and the smart contract returned back the results shown in Figure 7.



```
[ retrieveDataLakeUsingSHA256(bytes32) method Response ]
 » uint256 : 0
 » uint256 : 0
 » address : 0x6c56618BCbF502b237369551cF2Af7317E763eDb
 » bytes32 : 0x5a221f47e54beac4c9548116bf9196a23b041c0c664280d018c96a1089643568
 » uint256 : 1648624888
```

Figure 7 The SHA 256 matching

As depicted in Figure 7, the SHA256 value of the given DL matches the one of the specific DL that was successfully registered into the contract. In case the DL was compromised, the SHA256 value would not match the one on the smart contract and, therefore, the end-user would be directed not to execute any action using the specific DL. Table 2 presents the minimum time required to call the core functions of the smart contract.

Table 2 Minimum time of calling the core functions

Core Functions	**Time(s)**
registerDataLake()	≈10
retrieveDataLakeUsingSHA256()	≈1ms

As can be observed in Tables 1 and 2, the cost and the time required to call the main functions of the smart contract are not prohibitive.

V. CONCLUSIONS

This paper proposed a novel framework for standardizing the processes of storing/retrieving IoT data combined with data generated by heterogeneous sources to/from a DL organized with ponds architecture and focusing on providing security and privacy. The framework is based on a metadata semantic enrichment mechanism which uses the notion of blueprints to produce and organize specific meta-information called (Data Lake Blueprint (DLB), which is related to each source that produces data to be hosted in a DL. In this context, each data source is described via two types of blueprints which essentially utilize the 5Vs Big Data characteristics Volume, Velocity, Variety, Veracity and Value: The first includes information that is stable over time, such as, the name of the source and its velocity of data production. The second involves descriptors that vary as data is produced by the source in the course of time, such as the volume and date/time of production.

The goal of the framework presented in this work is to ensure that the DLB metadata will not be altered or modified. To this end, a Blockchain smart contract was developed aiming at providing the ability to a DL owner to register the DL into the Blockchain based on its metadata, and at allowing end-users to verify the correctness of a DL before conducting any actions on it.

Each time a new version of a DLB metadata is created a new SHA256 value for that specific DLB is generated automatically and is used along with the DL's *id* to register the specific DL into the Blockchain via the proposed smart contract. Multiple versions of a DLB can be created for each DL that results in the creation of a unique DL Chain for each DL and a whole new Chain for the full system, namely, DLMetaChain.

To demonstrate the effectiveness and applicability of the proposed framework the proposed smart contract was deployed and assessed on the Ethereum Rinkeby Test Network with very positive results.

Future work will include the investigation of how the proposed framework will integrate a new architectural style that is considered the evolution of DLs, namely data mesh. Further to that, investing more on one of the main challenges of DLs, that is, Security, Privacy and Data Governance, NFT blockchain technologies will be embedded to allow for sharing data stored in DLs with verified owners of this information. Finally, the concept of Process Mining (PM) methodologies will be incorporated, where a process is modelled not only using log files data but any structured, semi-structured and unstructured data from different data sources present in DLs, something which will allow extending existing PM algorithms and techniques.

Acknowledgment

This paper is part of the outcomes of the Twinning project DESTINI. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 945357.

References

[1] P. Sethi and S. R. Sarangi, "Internet of Things: Architectures, Protocols, and Applications," J. Electr. Comput. Eng., vol. 2017, 2017.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] P. Barnaghi, A. Sheth, and C. Henson, "From data to actionable knowledge: Big data challenges in the web of things," IEEE Intell. Syst., vol. 28, no. 6, 2013.K. Elissa, "Title of paper if known," unpublished.

[3] "How much data is created every day? [27 powerful stats]," SeedScientific, 07-Feb-2022. [Online]. Available: https://seedscientific.com/how-much-data-is-created-every-day/. [Accessed: 20-Mar-2022]. .

[4] P. P. Khine and Z. S. Wang, "Data lake: a new ideology in big data era," ITM Web Conf., vol. 17, p. 03025, 2018.

[5] P. Sawadogo and J. Darmont, "On data lake architectures and metadata management," J. Intell. Inf. Syst., 2020.

[6] P. Christodoulou, A. S. Andreou, and Z. Zinonos, "Skillschain: A decentralized application that uses educational robotics and blockchain to disrupt the educational process," Sensors, vol. 21, no. 18, pp. 1–9, 2021.

[7] B. M. Balachandran and S. Prasad, "Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Business Intelligence," Procedia Comput. Sci., vol. 112, pp. 1112–1122, 2017.

[8] M. J. Cafarella, A. Halevy, and N. Khoussainova, "Data integration for the relational web," Proc. VLDB Endow., vol. 2, no. 1, pp. 1090–1101, 2009.

[9] O. Hassanzadeh, K. Q. Pu, J. Miller, L. Popa, and M. A. Hern, "P445-Hassanzadeh.Pdf," vol. 6, no. 6, pp. 445–456, 2013.

[10] Y. Roh, G. Heo, and S. E. Whang, "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," IEEE Trans. Knowl. Data Eng., vol. PP, no. c, pp. 1–1, 2019.

[11] J. Fan, M. Lu, B. C. Ooi, W. C. Tan, and M. Zhang, "A hybrid machine-crowdsourcing system for matching web tables," Proc. - Int. Conf. Data Eng., pp. 976–987, 2014.

[12] P. P. Khine and Z. S. Wang, "Data lake: a new ideology in big data era," ITM Web Conf., vol. 17, p. 03025, 2018.

[13] M.Pingos and A. Andreou, "A Data Lake Metadata Enrichment Mechanism via Semantic Blueprints", In Proceedings of the 17th International Conference on Evaluation of Novel Approaches to Software Engineering , ISBN 978-989-758-568-5, ISSN 2184-4895, pages 186-196.

[14] P. N. Sawadogo, É. Scholly, C. Favre, É. Ferey, S. Loudcher, and J. Darmont, "Metadata Systems for Data Lakes: Models and Features," in Communications in Computer and Information Science, 2019, vol. 1064, pp. 440–451.Sdfsdfsd

[15] A. Beheshti, B. Benatallah, R. Nouri, and A. Tabebordbar, "CoreKG: A Knowledge lake service," Proc. VLDB Endow., vol. 11, no. 12, pp. 1942–1945, 2018.

[16] L. A. Linn and M. B. Koo, "Blockchain For Health Data and Its Potential Use in Health IT and Health Care Related Research," ONC/NIST Use Blockchain Healthc. Res. Work., pp. 1–10, 2016.

[17] H. F. Atlam, A. Alenezi, M. O. Alassafi, and G. B. Wills, "Blockchain with Internet of Things: Benefits, challenges, and future directions," Int. J. Intell. Syst. Appl., vol. 10, no. 6, pp. 40–48, 2018.

[18] Z. Zinonos, P. Christodoulou, A. Andreou, and S. Chatzichristofis, "ParkChain: An IoT parking service based on blockchain," Proc. - 15th Annu. Int. Conf. Distrib. Comput. Sens. Syst. DCOSS 2019, pp. 687–693, 2019.

[19] P. Raj, *Empowering digital twins with blockchain*, 1st ed., vol. 121. Elsevier Inc., 2021.

[20] M. P. Papazoglou and A. Elgammal, "The manufacturing blueprint environment: Bringing intelligence into manufacturing," 2017 Int. Conf. Eng. Technol. Innov. Eng. Technol. Innov. Manag. Beyond 2020 New Challenges, New Approaches, ICE/ITMC 2017 - Proc., vol. 2018-Janua, pp. 750–759, 2018.

[21] F. Casino, T. K. Dasaklis, and C. Patsakis, "A systematic literature review of blockchain-based applications: Current status, classification and open issues," Telemat. Informatics, vol. 36, no. May 2018, pp. 55–81, 2019.

[22] J. Moreno, E. Fernandez-Medina, E. B. Fernandez, and M. A. Serrano, "BlockBD: A security pattern to incorporate blockchain in big data ecosystems," ACM Int. Conf. Proceeding Ser., 2019.