Security and Ownership in User Defined Data Meshes

Michalis Pingos *, Panayiotis Christodoulou , Andreas S. Andreou *

Posted Date: 5 March 2024

doi: 10.20944/preprints202403.0265.v1

Keywords: Big Data; Smart Data Processing; Systems of Deep Insight; Data Meshes; Data Lakes; Data Products; Blockchain; NFT; Data Blueprints

 doi:10.20944/preprints202403.0265.v1

Article

Security and Ownership in User Defined Data Meshes

Michalis Pingos [1], Panayiotis Christodoulou [2] and Andreas S. Andreou [1]

[1] Cyprus University of Technology; michalis.pingos@cut.ac.cy; andreas.andreou@cut.ac.cy
[2] University of Nicosia; christodoulou.pa@unic.ac.cy
* Correspondence: michalis.pingos@cut.ac.cy

Abstract: Data Meshes is an approach to data architecture and organization that treats data as a product and focuses on decentralizing data ownership and access. It has recently emerged as a field that presents quite a few challenges related to data ownership, governance, security, monitoring, and observability. To address these challenges, this paper introduces an innovative algorithmic framework leveraging data blueprints to enable the dynamic creation of Data Meshes and Data Products in response to user requests, ensuring that stakeholders will have access to specific portions of the Data Mesh as needed. Ownership and governance concerns are addressed through a unique mechanism involving Blockchain and Non-Fungible Tokens (NFTs). This facilitates secure and transparent transfer of data ownership, with the ability to mint time-based NFTs. By combining these advancements with the fundamental tenets of Data Meshes, this research offers a comprehensive solution to the challenges surrounding data ownership and governance. It empowers stakeholders to navigate the complexities of data management within a decentralized architecture, ensuring a secure, efficient, and user-centric approach to data utilization. The proposed framework is demonstrated using real-world data from a poultry meat production factory.

Keywords: big data; smart data processing; systems of deep insight; data meshes; data lakes; data products; blockchain; NFT; data blueprints

1. Introduction

Nowadays, Big Data can be characterized as the «new oil» as it is recognized as a valuable human asset. Effective aggregation and analysis of this data may unearth information that provides insights into numerous facets of everyday activities and offers the ability to anticipate future occurrences. Big Data refers to the substantial volumes of digital information consistently produced by machine and global population from diverse sources such as social media, Internet of Things (IoT) devices, machines and sensors logs, public records and open data, online transactions, websites and applications, research and scientific instruments, etc. [1]. The vast majority of Big Data originates from heterogeneous data sources, yielding a variety of data types that include structured, unstructured, and semi-structured data. Encompassing a diverse range of content, Big Data spans from textual information to multimedia elements, such as images, videos, and audio [2].

The three primary characteristics (3Vs) of Big Data, as presented by Dough Laney in 2001, form and define its fundamental framework [3]. Firstly, *Volume* represents the broad amount of data generated from data sources, often reaching high levels that challenge typical data processing methods. The second characteristic defining the tempo with which data is created, processed, and made available for analysis is denoted by *Velocity*. Fast processing speeds are required to keep up with the increasing rate of data creation due to the emergence of real-time data sources like social media and sensors. Thirdly, the term *Variety* highlights the variety of data kinds, encompassing organized, unstructured, and semi-structured information. By integrating a broad range of textual, visual, and audio information, this inclusivity recognizes that Big Data extends beyond traditional databases. Taken together, these three qualities create the foundation for realizing and capitalizing on the possibilities of Big Data in a data-driven modern world. In addition, seven more characteristics were included to this list after 2001 leading to the 10Vs term for Big Data. The new properties are

 doi:10.20944/preprints202403.0265.v1

Value, *Veracity*, *Volatility*, *Validity*, *Vulnerability*, *Variability*, and *Visualization* [4] and offer additional descriptive assets of Big Data.

In the pre-Big Data era storage designs were mostly based on file systems and conventional relational databases. Relational databases with clear schemas, like MySQL and Oracle, were great at handling structured data. A lot of people used file-based storage systems, such us Network Attached Storage (NAS) and Storage Area Network (SAN), to store documents and other kinds of files. During the same period, the conventional approach to address escalating data requirements involved vertical scaling, which entailed augmenting resources on a single server [5]. In the era of Big Data, which is characterized by immense data volumes, rapid data transfer rates and the diversity of weakly structured data from numerous heterogeneous sources, as declared also by the 10Vs characteristics, resulted in a fundamental transformation of storage architectures. NoSQL databases, such as MongoDB and Cassandra, as well as distributed storage systems like Hadoop Distributed File System (HDFS), have now become more popular [6].

The complex interactions amongst Data Lakes, Data Meshes, and Data Markets in the Big Data era create a dynamic ecosystem that transforms how businesses manage and extract value from heterogeneous data sources and Big Data [7]. Data Meshes and Data Markets are innovative data management frameworks introduced in 2019 by Zhamak Dehghani, diverging from the conventional approach of Data Lakes storage architectures. These storage architectures and structures can be deployed with storage and processing technologies, such as Apache Hadoop, Apache Spark, or cloud-based solutions like Amazon S3, Azure Data Lake Storage, or Google Cloud Storage [8]. While these frameworks are linked with Big Data Processing, the primary unsolved challenging problems revolve around security, encompassing issues related to privacy, regulatory requirements, and access control. Notably, weaknesses in metadata management pose challenges, as data in lakes or meshes can be replaced without proper oversight of the contents [9].

The primary research contribution of this paper lies in the introduction of an innovative framework that leverages Semantic Data Blueprints (SDB) [10] for the dynamic assembly of Data Meshes and data products responding to user demands on one hand, and ensuring that stakeholders access specific areas of the Data Mesh as needed via transfer of ownership on the other. The integration of non-fungible tokens (NFTs) and Blockchain technology collaboratively establishes a novel approach to address data ownership and governance concerns. The core of the framework is a dedicated algorithm which involves the execution of specifics steps to facilitate secure and transparent data ownership transfers by incorporating the ability to mint time-based NFTs with extended functionality.

The proposed approach builds upon and expands earlier research on the subject that proposed SDB, a semantic metadata enrichment technique for Data Lakes that enables the effective storing and retrieval of data from distributed and heterogeneous data sources and ensuring security in Data Lakes using Blockchain technology and NFTs [10,11,12]. The same concepts are employed in this work but this time they align with the characteristics of Data Meshes, ensuring security and ownership through the integration of Blockchain and NFT technology, thereby paving the way for the development of Data Markets. In this context, a Data Mesh is thought of as the evolution of a Data Lake in terms of managing massive amounts of data (Big Data) expressed in a variety of formats (structured, unstructured, and semi-structured), but most crucially, for making it simple, rapid, and effective to trace.

Real-world manufacturing data from Paradisiotis Group (PARG), a significant local industrial player in Cyprus, is used to illustrate the proposed approach. PARG is one of the most significant companies and experts in the field of poultry farming and production/trading of poultry meat in Cyprus. It provides a large assortment of food products which are delivered to local supermarkets. The operational procedures and production data of the factory are treated as confidential for privacy and security reasons. Consequently, this work uses a masked and de-identified rendition of the data and only presents a portion of the processes, providing limited but specific details. However, the case study reported in the present paper successfully illustrates the fundamental ideas of the proposed framework, confirming its applicability and effectiveness.

 doi:10.20944/preprints202403.0265.v1

The remainder of the paper is structured as follows: Section 2 discusses the technical background and related work in the areas of Data Lakes and Data Meshes. Section 3 outlines previous work performed on the semantic enrichment technique, which is adopted and extended in this work to address security and ownership aspects. Section 4 presents the extended Data Meshes framework and discusses its main components. This is followed by demonstrating the applicability and assessing the performance of the proposed framework in Section 5 through a case-study conducted using real-world data collected at PARG. Finally, Section 6 concludes the paper and highlights future research directions.

2. Technical Background

2.1. Understanding Data Lakes and Data Meshes

A Data Lake (DL) is a centralized architecture designed to store vast amounts of structured, unstructured and semi-structured data at any scale. Unlike traditional databases or data warehouses that require data to be structured before storage, a DL allows for hosting raw data in its native format [13]. This means data from various sources like logs, clickstreams, social media, videos, and sensor data can be stored without the need for pre-defined schemas. DLs offer storage flexibility, allowing the storage of data in its raw form without upfront schema definition. This feature enables the accommodation of various data types and formats from diverse sources at any production frequency. DLs are highly scalable and capable of handling very large datasets making them ideal for Big Data applications. They often provide cost-effective storage options by leveraging cloud object storage, resulting in more economical solutions compared to traditional data warehouses.

DLs seamlessly integrate with tools and technologies that enable processing, querying, and analyzing the stored data. Properly configured DLs can implement security measures and data governance policies to ensure privacy and compliance with regulations. While DLs offer a high degree of flexibility, they require careful management to prevent them from becoming "data swamps", that is, hosting places where data is poorly organized, difficult to find, and hard to analyze [14]. To address this concern, practices like metadata management, data cataloguing, and establishment of data governance policies are crucial. Figure 1 presents the structure of a DL and an algorithmic description of how the DL concept works in practice, from collecting the data, annotating it using metadata, storing it and finally retrieving it based on the metadata tags.

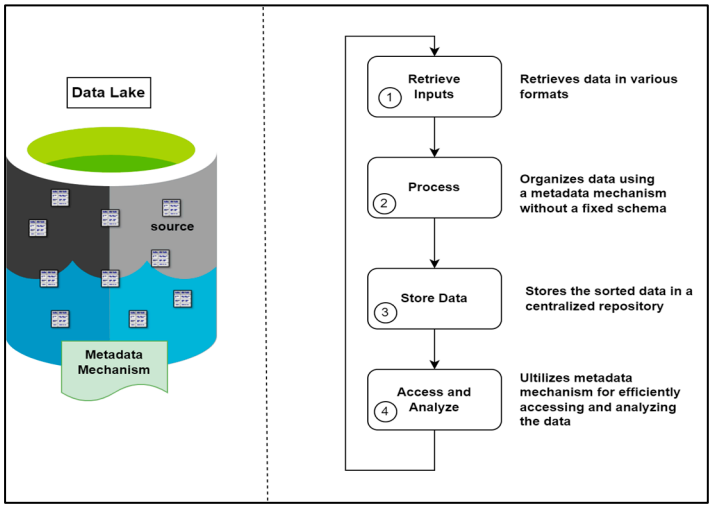

Figure 1. Data Lake architecture and the concept algorithmic approach.

The concept of Data Mesh (DM), as mentioned in the previous section, was introduced in 2019 [15], which essentially represents a novel approach to data management within large organizations. Unlike traditional methods, a DM emphasizes several key concepts to revolutionize data handling. Firstly, it advocates for Domain-oriented Ownership. This means that data domains are entrusted to

 doi:10.20944/preprints202403.0265.v1

the teams or business units possessing the highest expertise in that specific domain. These teams bear the responsibility for ensuring the quality, accessibility, and privacy of their respective domain's data. Additionally, a DM promotes the idea of Decentralized Data Products. Here data is treated as a product and each domain team is accountable for the entire data lifecycle within their domain. This encompasses tasks such as production, consumption, quality assurance, privacy measures, and comprehensive documentation. Furthermore, DMs advocate for Federated Computational Governance, an approach where each domain team defines and enforces the computational logic specific to their domain. This logic is then executed within the broader context of the mesh [16].

To facilitate autonomy and efficiency, DM incorporates a self-serve data infrastructure. This infrastructure is designed to empower domain teams with the necessary tools and resources to independently manage their data products, reducing reliance on centralized data engineering teams. Embracing an API-first approach, DM encourages the utilization of Application Programming Interfaces (APIs) for seamless data exchange and communication between different components of the system. This promotes loose coupling and flexibility in how data is consumed and utilized. Figure 2 presents the structure of a DM and the algorithm that serves as the core of its operation.

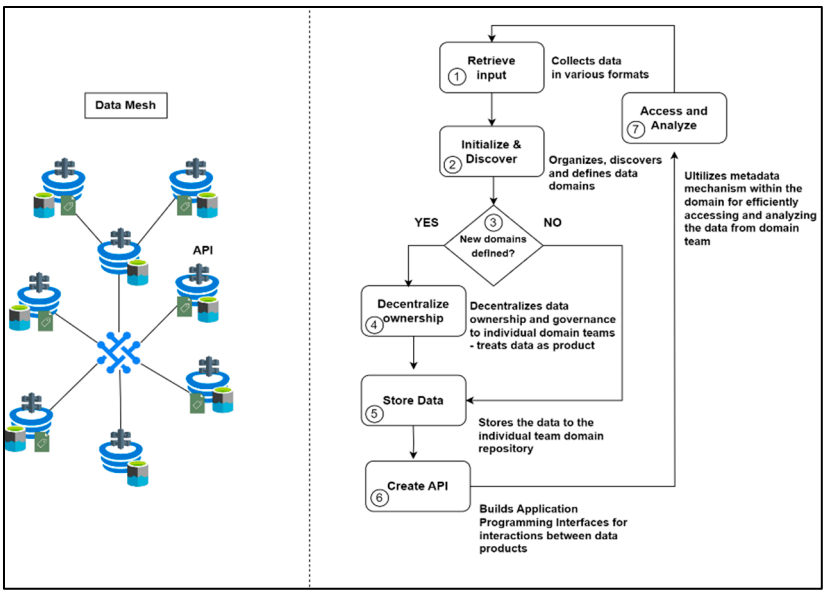

Figure 2. Data Mesh architecture and the concept algorithmic approach.

Furthermore, DM emphasizes a holistic view of the data product lifecycle. This encompasses stages such as discovery, ingestion, processing, storage, access, and consumption. Each of these stages is to be carefully considered and managed by the respective domain teams, ensuring a comprehensive and efficient data handling process. By adopting a DM approach, organizations aim to address the challenges of scaling data operations in a complex environment, where multiple teams work on diverse data domains. It provides a framework for decentralizing data ownership and enabling more effective, scalable, and resilient data operations.

Conversely, a Data Market is an ecosystem or marketplace where individuals, companies, or systems can buy, sell, or exchange data by leveraging the idea of DMs. Data suppliers in a Data Market offer datasets for purchase or access by data consumers for a range of applications, such as analysis, research, machine learning, and more [7]. Data Markets facilitate the efficient sharing and monetization of data, allowing businesses to leverage external sources of information to enhance their insights and decision-making processes.

Using a large manufacturing company in the field of poultry farming and poultry meat trading as our case-study and example demonstrator, we were able to identify various operational areas, including livestock records, agricultural data, supply chain management information, financial transactions, and trading analytics. Each operational area is assigned to a specialized team responsible for its monitoring and upkeep. Moreover, each team is tasked with generating specific

 doi:10.20944/preprints202403.0265.v1

data products tailored to their respective domains, such as APIs for accessing data, algorithms for analyzing trading trends, tools for secure data sharing, and reporting mechanisms for financial analytics. This makes the proposed approach an ideal way of sharing portions of data across authorized groups (e.g. departments) or individuals. Adopting a federated computational governance approach ensures that each team defines and enforces the computational logic for their specific domain, facilitating the implementation of specialized algorithms and quality checks. Additionally, each team has access to a self-serve data infrastructure, equipped with tools and resources for managing their data products independently, thereby ensuring autonomy and operational efficiency. To enhance interoperability within the field, the implementation of APIs and adherence to industry standards are prioritized, allowing seamless communication and data exchange between different operational areas. This approach contributes to the optimization of data management and the creation of tailored products, ultimately benefiting stakeholders in poultry and farming trading, including producers, traders, and administrators.

It is important to note that while DM is more about organizational and conceptual principles for data management, DL refers specifically to the technology and infrastructure for storing large volumes of raw data. These concepts are not mutually exclusive, and, in practice, organizations can implement a DM framework while utilizing a DL as the underlying basic component of their technical infrastructure for data storage and processing.

2.2. Understanding Blockchain and NFTs

Blockchain serves the purpose of providing secure and transparent means for recording and transferring data. Notably, it addresses privacy concerns by anonymizing personal data, contributing to its increasing popularity and integration into infrastructure, opening avenues for innovative applications [17]. Functioning as a decentralized database on a peer-to-peer network, Blockchain establishes a distributed communication network enabling non-trusting nodes to interact without relying on a central authority. Its protocols ensure a verifiable and trustworthy system, offering traceability, transparency, and enhanced privacy and security features. In essence, Blockchain is evolving into a fundamental technology with wide-ranging applications and use-cases such as IoT, Smart Contracts, NFTs, Cybersecurity and Cryptocurrency, providing a foundation for secure and trustworthy data transactions [18].

Algorithmically, Blockchain includes a number of essential elements, procedures, and guidelines to create a strong and feature-rich decentralized system. Initializing basic elements, such as a consensus mechanism and cryptographic algorithms for secure key management and hashing, are the first steps in the process. Implementing token and smart contract standards like ERC-20 and ERC-721 increase functionality by managing the creation, transfer, and ownership of assets [19]. With zero-knowledge proofs as an example, the method smoothly incorporates Decentralized Identity Standards (DIDs) to guarantee secure identification and privacy standards, offering strong user data security. Interledger Protocol and other interoperability standards also make cross-chain communication easier [20]. The integration of decentralized storage protocols, such IPFS, ensures file storage that is dispersed and impervious to censorship. Governance norms support secure and efficient decision-making. Security measures provide protection against vulnerabilities, compliance standards guarantee conformity to legal requirements, and governance standards support efficient decision-making. This all-encompassing strategy creates a conceptual framework for the building of a Blockchain that integrates fundamental criteria, promoting a safe, compatible, and considerate decentralized ecosystem.

Non-Fungible Tokens (NFTs) are a ground-breaking innovation in the ownership and management of digital assets. Because every NFT is distinct and has a unique identifier, it cannot be copied or traded. Blockchain technology is used to accomplish this uniqueness. NFTs are used to verify ownership of a wide range of digital and physical goods [21], including digital art, music videos, real estate, gaming avatars etc. NFTs are also crucial to Web 3.0, the next iteration of the Internet that many companies and analysts are pushing. Blockchain's decentralized structure

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2024 doi:10.20944/preprints202403.0265.v1

guarantees the integrity and transparency of ownership data, and smart contracts streamline transactions by automating tasks like ownership transfers and royalty distribution.

Finally, the NFT process algorithm starts with the digital asset being initialized, having its nature defined, and being given a unique identification. The implementation of a smart contract that oversees the NFT requires integration with a Blockchain platform, such as Ethereum, via the ERC-721 standard [22]. The NFT is created during the minting process by adding ownership information and other pertinent metadata to the smart contract. Smart contract updates enable ownership transfers, guaranteeing safe and transparent transactions documented on Blockchain. The NFT ecosystem is made more efficient by automating features in the smart contract, such as the distribution of royalties upon resale. NFTs are posted on NFT marketplaces such as OpenSea or Rarible, where buyers and sellers can transact to make them more widely available [21]. Verifying the integrity of related metadata and examining ownership records on the Blockchain are two steps in the process of authenticating NFTs. The foundation of the NFT lifecycle is the aforementioned algorithmic procedure, which provides a methodical way to create, transfer, and confirm ownership of distinct digital assets on the Blockchain. The whole process is depicted graphically in Figure 3.

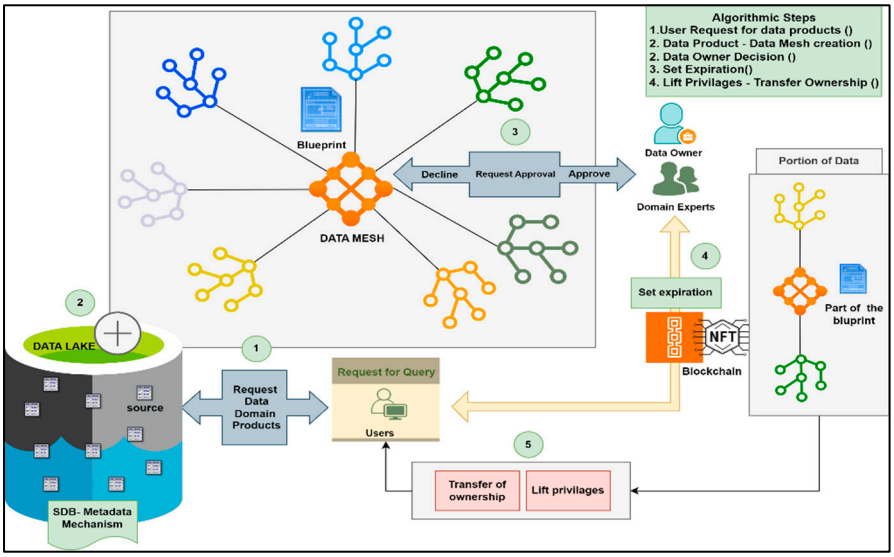

Figure 3. The algorithmic process of transferring ownership from the data owner to a user.

3. Related Work

The combination of DLs, DM, and Blockchain-based technologies—specifically, NFTs—in the field of modern data management creates a dynamic synergy that is changing how businesses handle data ownership, accessibility and storage. DLs function as large storage spaces for heterogeneous data, promoting a single repository that can handle a variety of data types. In addition, the DM paradigm supports distributed data processing and domain-oriented ownership using decentralised data architectures. A new dimension is brought to data ownership and authenticity by the integration of NFTs on Blockchain platforms, which offer a safe and verifiable framework for identifying the provenance and ownership of individual pieces of data. The integration of Blockchain, DLs and DMs improves the scalability and flexibility of data ecosystems and lays the groundwork for more open, safe, and cooperative data management procedures as related work unveils.

A dedicated DL architecture was used in [13] to investigate how Blockchain technology might be integrated to improve the scheme's metadata management. It specifically presented the use of NFTs that are stored on the Blockchain to represent metadata for every data source. The intention was to use Blockchain technology to improve the DL's semantics metadata, which could lead to better data management, organization, and retrieval. Furthermore, [11] addressed the challenges associated with smart processing of Big Data in the context of DLs, as well as ownership and security using Blockchain and NFT technologies. It emphasized the need for a disciplined approach to manage diverse data sources within DLs for predictive and prescriptive analytics. That paper introduced a

 doi:10.20944/preprints202403.0265.v1

novel standardization framework that integrates the 5Vs of Big Data characteristics and blueprint ontologies. The framework utilized a ponds architecture to organize DLs and incorporated a metadata semantic enrichment mechanism for efficient storage and retrieval. Notably, the mechanism supported visual querying and enhanced security through Blockchain and NFTs. The authors also provided a comparative analysis with other metadata systems, demonstrating promising results based on a set of functional properties.

An enhanced DL metadata framework called DLMetaChain was introduced in [12], which can manage data from diverse sources like IoT data using Blockchain. The paper discussed the changing IoT ecosystem, where a variety of sources produce large amounts of data that are then converted into useful information. Metadata management becomes difficult when storing such data, including IoT data, in repositories like DLs, especially when it comes to security and access control. The principal aim was to design an architecture that utilizes Blockchain technology to guarantee the data integrity of the DL by impeding any unsanctioned changes or additions.

A visionary approach to establish a distributed federated medical DL and ecosystem was proposed in [23], involving hospitals and personal health data from wearable medical devices. It emphasized the creation of a Blockchain-based platform with commercial incentives, addressing data ownership, patient privacy, and controlled access. The platform facilitated owner-centric medical data exchange, securely aggregated data from various hospitals, and unlocked academic and business value by representing medical data as NFTs. The primary goal was to improve healthcare research while fostering a sustainable medical data ecosystem.

Finally, in order to manage data at scale, [24] investigated how a Blockchain-powered metadata catalogue might be integrated into a DM architecture. The metadata catalogue improved governance, efficiency, access, and discovery. The catalogue managed metadata across a dispersed network of data domains with federated governance, immutability, and transparency thanks to the use of Blockchain technology. A proof-of-concept solution utilizing HyperLedger Fabric was presented, with advantages including increased reliability, efficiency, and transparency being highlighted. It also discussed and suggested possible solutions for issues including governance, scalability, and interoperability.

4. A Framework for Supporting Transfer of Ownership in Data Meshes

This section describes the proposed framework for transferring ownership of data products residing in DM. The framework follows a series of algorithmic steps that include the creation of the DM through its transformation from a DL that bears a specific architectural structure, and the development of the appropriate smart contracts the execution of which facilitates the transfer and proves ownership of a specific data product.

4.1. Semantically Enrisched Data Lake Architecture and Data Mesh Products Creation

A metadata mechanism is of paramount importance for a DL as it functions as its organizational backbone, offering a systematic and detailed catalog of the diverse datasets hosted within the DL. Without such a metadata mechanism a DL will gradually be transformed into a "Data Swamp". In essence, a metadata mechanism provides data owners with a vital insight into the type and context of the stored information by capturing important details about the origin, structure, relationships, and usage of data. By providing this information navigating and mapping the raw data becomes feasible, something that makes data search, retrieval, and management easier and efficient.

The SDB is a metadata enrichment mechanism that identifies and characterizes a candidate source before it becomes member of a DL [10]. The framework described in [10] integrates blueprint ontologies with the 5Vs Big Data features, namely Volume, Velocity, Variety, Veracity, and Value, to support data processing (storage and retrieval) in DLs organized with a pond architecture. The latter structures a DL in several distinct data ponds, each of which holds or refers to a certain type of data according to the pond design. Depending on the type of data (structured, semi-structured, unstructured), each pond has a unique data processing and storage method. When extracting data

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2024 doi:10.20944/preprints202403.0265.v1

from the DL, this built-in pond architecture is quite useful as it supports quick and easy access to the storage space.

As previously mentioned, a dedicated blueprint is developed to describe each data source storing data in the DL. Specifically, the blueprint of a source consists of two interconnected blueprints as shown in Figure 4, the stable and the dynamic blueprint [10]. The former is static and describes the name and type of the source, the type of data it produces, as well as the value, velocity, variety, and veracity of the data source pushed in the DL. The latter is a dynamic blueprint which involves attributes that are not stable over time and essentially characterizes volatile properties such as the volume of data, the last source update, and keywords characterizing the source. The dynamic blueprint is updated every time data sources produce new real-time or batch data, or its description through keywords may be modified. In essence, the metadata description - SDB is provided in Terse Triple Language (TTL) using the Resource Description Framework (RDF), which is a well-known framework for describing resources on the Web. The metadata mechanism contains TTL descriptions for all the sources included in the DL.

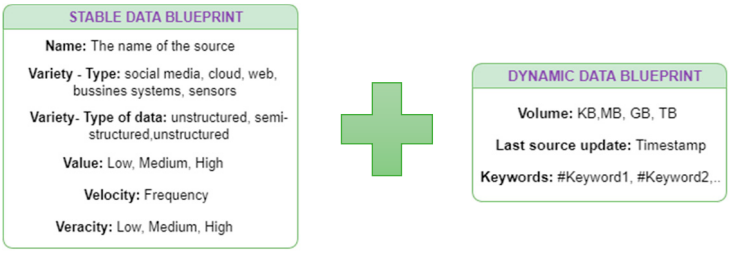

Figure 4. Data source blueprints description using 5Vs of Big Data.

In essence, TTL is a serialization format that provides a concise and human-readable way to represent RDF data, making it easier for both machines and humans to work with semantic information on the Web. RDF represents information as triples, which consist of subject-predicate-object statements. The resource being described is the subject, the property or attribute is the predicate, and another resource or value is the object. An example of a triple may be *ex:variety "unstructured"*, which means that the subject is the source, the predicate is "variety" and the object is the value "unstructured".

Let us assume that a user requests access to specific sources producing data and storing it to the DL. In this case a dedicated SPARQL query is formed and executed on the DL. When the query starts executing, it first asks the owner of the data for her/his approval. If the owner approves the query, then the framework, and specifically the metadata mechanism of the DL, is utilized to create the corresponding DM data product that satisfies the query as presented earlier in Figure 3. Figure 3 also shows that the user has access only to the sources requested through the corresponding APIs. Furthermore, this access is restricted to the specific person and is valid only within a specific period via Blockchain and NFT technologies as will be presented with details in the next subsection.

Analytically, the steps taken are as follows:

1. The owner of the contract can add an administrator on the contract by calling the *addAmin()* function inserting an EVM-compatible address. Once an administrator is created, (s)he gets access through her/his address to certain admin-only functions on the contract.
2. An administrator can mint an NFT by executing the *safeMint()* function providing the address of the recipient, an expiration date in UNIX epoch time, the query that is associated with the NFT and its access level. If the value of the access level is set to 1, then the NFT grants read-only access to its new owner and the NFT is non-transferable, while, if the value is set to 2 the owner of the NFT, besides the read access, gets also transfer access and therefore can transfer the ownership of the NFT to a different user. At any given time, the current owner of the NFT can access and read the data.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2024 doi:10.20944/preprints202403.0265.v1

4.2. Smart Contract Architecture

The Blockchain-based architecture uses a specially designed ERC721 smart contract that was implemented to evaluate the use of the proposed framework. ERC721 is a standard that is used in EVM-compatible Blockchain networks to represent ownership of NFTs, where each token is unique and has its own metadata. In this work we decided to develop our smart contract based on the ERC721 standard for two main reasons: (i) with ERC721, users can securely own, transfer, and manage their digital assets with transparent and verifiable ownership records, and, (ii) the ERC721 standard ensures that NFTs can easily interact with several wallets and decentralized applications (dApps), enhancing their utility and accessibility.

The purpose of the smart contract developed in this paper is threefold: (i) Allow data owners to mint time-based NFTs and transfer them to an address; (ii) Allow NFT owners to read specific portions of data for a certain period of time; and, (iii) Allow NFT owners to transfer ownership of the data to a different user. The proposed smart contract consists of three main actors: The contract owner, who is the deployer of the contract and is also responsible for registering administrators onto the contract; the contract administrators, who oversee the minting process; the authorized users who can view or transfer data. The administrator algorithmic workflow of the proposed framework is depicted in Figure 5 and summarized in pseudocode.

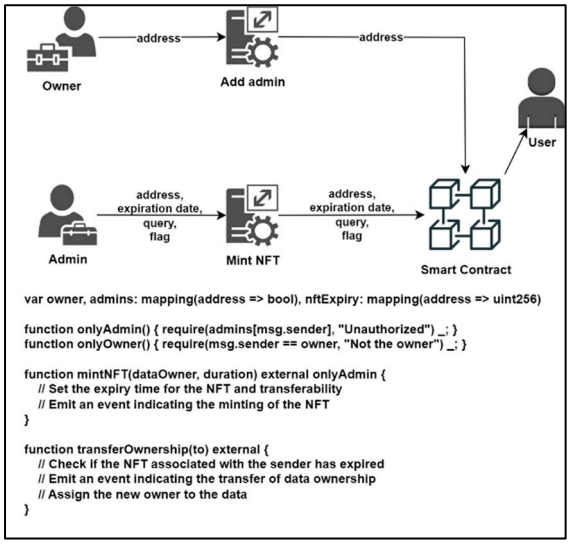

Figure 5. The admin algorithmic workflow and pseudocode.

Figure 6 presents the algorithmic workflow followed for authorized users. Authorized users can view the assigned data based on two parameters, the expiration date and the query. A user holding a valid NFT can access a token-gated website to view the data. The website checks the eligibility of the connected address to allow or refuse access to the user. Finally, as depicted in Figure 7, NFTs are separated into two categories, transferable and non-transferable. When an NFT is minted, the admin specifies if the token has read-only or transfer access. When a user who holds a specific non-expired NFT initiates a transfer function, the contract checks whether the token can be transferred or not to a different address and proceeds to accepting or rejecting the request accordingly. If the NFT is successfully transferred, then the new owner of the NFT is automatically granted access to the token-gated website and can view the data.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2024 doi:10.20944/preprints202403.0265.v1

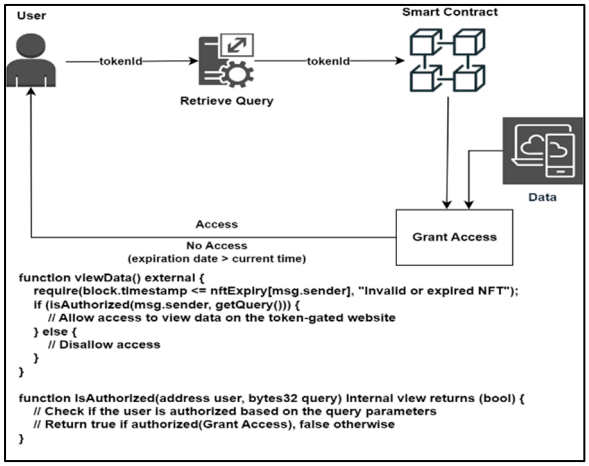

Figure 6. The authorized user algorithmic workflow and pseudocode.

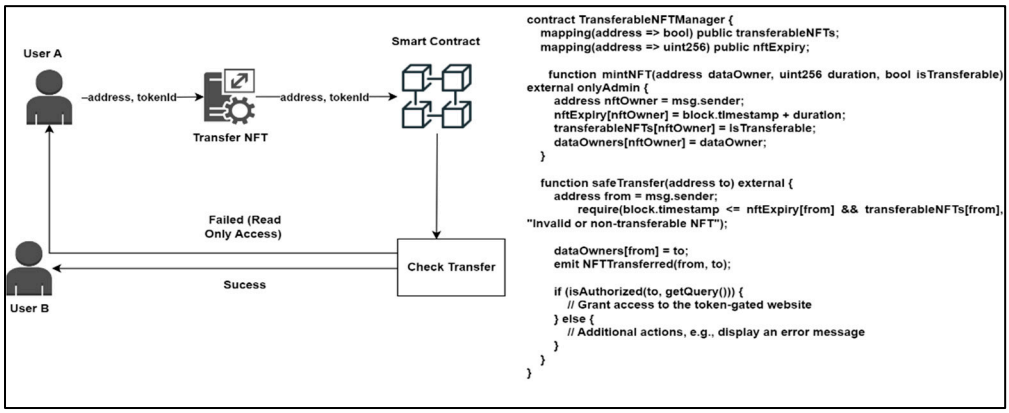

Figure 7. The algorithmic user workflow for transferable and non-transferable NFTs.

5. Framework Demonstration Through a Real-World Case-Study

5.1. The PARADISIOTIS Group (PARG) Factory Case-Study

As previously mentioned, this work utilizes a real-world case-study from the area of smart manufacturing to demonstrate the applicability of the framework. Specifically, it utilizes data recorded at the PARG factory, the main business line of which is chicken farming and poultry meat production and distribution. PARG is a continuously growing company that invested over the years in modern and technologically advanced equipment for the breeding processes (e.g., automatic ventilation system, technology assisted mill for mixing ingredients and preparing chick food, etc.) and the production line (cutting, mixing and packaging of poultry meat). The management of the factory constantly seeks to improve performance and quality levels by frequently adapting the production processes and adopting new technologies.

Data is produced within the factory mainly by two systems: (i) CUBORA is a fully operational heating control system designed to produce and monitor data related to poultry heating and emissions into the feeding atmosphere. This system is essential for ensuring the healthy growth and well-being of chicks on farms; and (ii) AGROLOGIC, which specializes in the field of automated climate controllers, feeding and weighting systems. AGROLOGIC is integrated with Chore Time controller and collects metrics from several remote sensors that are distributed into the farms, such as CO2, Temperature, Humidity, Air Static Pressure, and Light Intensity Level. All metrics are recorded in a database and are accessed through a Web application in real-time. Furthermore, images of the farms and/or equipment may be recorded for shift managers to inspect visually when

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2024 doi:10.20944/preprints202403.0265.v1

necessary. Finally, the system generates alerts if any of the metrics exceed pre-defined thresholds via an embedded GSM modem.

PARG case-study presents all characteristics of Big Data originating from heterogeneous sources with atypical patterns, which produce various kinds of structured, semi-structured, and unstructured data in high frequencies. This heterogeneous data needs to be treated differently than normal production speed data and be stored in more flexible and/or higher servicing speed data storage architectures or structures compared to classic Relational Databases and Data Warehouses, such as Big Data Warehouses, DLs and DMs. To this end, the current work developed a dedicated DL for PARG in a controlled (lab) environment and applied the basic principles of SDB, Blockchain and NFT technologies for creating Data Products and Domains. The latter are produced based on a DM constructed through the DL metadata mechanism. User requests for access to these data products are addressed to the Data Owner and then ownership may be granted through NFTs based on the relevant privileges, providing at the same time the ability to grant access and use the data only for a specific period of time.

5.2. Use-Case Scenarios

As previously mentioned, a request to access a DL is supported by the utilization of the SDB semantic enrichment mechanism, which is the cornerstone for creating a data product as part of the DM according to user preferences and ownership granting. Access and ownership for a specific period of time is recorded on Blockchain using a dedicated NFT. The use-case of the PARG factory focuses on the department of poultry feeding where sources produce data during the feed-cycle of chicken within a specific farm. An excerpt of the structure of the corresponding SDB is depicted in Figure 8. We have selected the following metadata characteristics to describe a source which produces data for monitoring the chicken flock farming in different locations: (i) Source Name; (ii) Location; (iii) Feed cycle start; (iv) Feed cycle end; (v) Keywords; (vi) Variety; (vii) Velocity; (viii) Volume, and, (ix) Source Path.

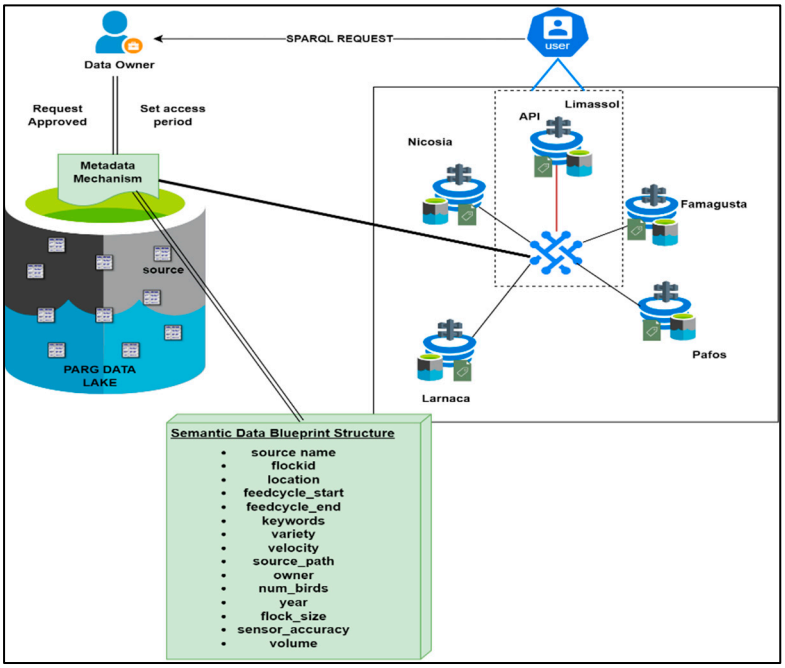

Figure 8. PARG's Data Lake SDB structure for the use-case scenarios.

The use-case scenarios tested are based on user requests to access a specific portion of data. For example, PARG stakeholders (shift managers, farm carers, production workers) often need to consult data related to the number of chicks in a farm, the environmental conditions within a farm, electricity consumption, emissions in the atmosphere, biomass production, etc. Therefore, data products in the

 doi:10.20944/preprints202403.0265.v1

DM were constructed to reflect these pieces of available information. Furthermore, in the scenarios below we assume also that the shift manager wishes to acquire access to all information related to the Limassol farm and that at some point (s)he wishes to transfer this access to the head of production. Normally, access permissions are requested by sending a message to the owner of the data through a dedicated SPARQL. This query is essentially executed in all scenarios that follow. Essentially, a user requests access to specific portion (data sources) of a DL through the DM data products that are constructed to provide information for location "Limassol" as presented also in Figure 8.

To evaluate the efficiency of the proposed framework we first developed a dedicated smart contract that was deployed on the Sepolia Test Network and then we executed a series of transactions. The smart contract's address is *0x88790ed3407e3b395ab0276d530 5a273a497612b* and the contract owner is *0xfb43d1384FC250B59996933CA2D8C766722 7BE52*. The reader can refer to the smart contract's URL on etherscan.io for complete access to the source code of the contract.

By using the smart contract, we have explored various scenarios to showcase its fundamental features that include NFT minting with additional on-chain information, data retrieval and transfer restrictions.

5.2.1. Scenario 1 - Minting

As previously mentioned, this scenario demonstrates how a user that wishes to access all sources of the factory that produce data during breeding with location the city of Limassol is serviced by executing the SPARQL query listed below. This scenario illustrates the minting capability of the smart contract, as outlined in the proposed administrative algorithmic workflow framework.

Initially, the admin of the smart contract with address *0xfb43d1384FC250 B59996933CA2D8C7667227BE52* executes sequentially two token processes for minting transactions to the address *0xcF1aB65AE4EFaA9BE8cDB13078360B811D11616D*, the first not allowing the token to be transferrable (i.e., the ownership of the data may not be passed on to another user) and the second allowing to do so. The processes are executed with the following parameters:

First token process parameters:

Date of Expiration: 1706094000 (Wednesday, 24 January 2024 11:00:00 UTC)

Query: PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ex: <http://example.org/>
SELECT ?location ?sourcePath
WHERE {
?source rdf:type ex:Description;
ex:location "Limassol";
}

Transferrable: NO (flag is set to 1)

Second token process parameters:

Date of Expiration: 1706095000 (Wednesday, 24 January 2024 11:16:40 UTC)

Query: PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ex: <http://example.org/>
SELECT ?location ?sourcePath
WHERE {
?source rdf:type ex:Description;
ex:location "Limassol";
}

Transferrable: YES (flag is set to 2)

Once the transactions are confirmed on the Blockchain network, the address *0xcF1a B65AE4EFaA9BE8cDB13078360B811D11616D* becomes the owner of both token ids #0 and #1, as depicted in Figure 9. Essentially, the owner has access to the PARG sources for Limassol's farm with

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2024 doi:10.20944/preprints202403.0265.v1

either token#0 or token#1. The main difference between the two tokens is the ability to use them for transferring ownership to another user as will be demonstrated below.

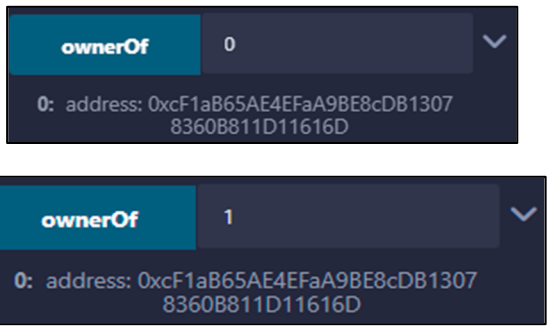

Figure 9. Owner for token ids #0 and #1.

5.2.2. Scenario 2 – Retrieving Data

This scenario presents the retrieving capabilities of the smart contract which are based on certain requirements. Here we are using address *0xcF1aB65AE4EfaA9BE8cDB13 078360B811D11616D* that corresponds to the owner of both NFTs #1 and #2. This address is checked to comply with two restrictions: First, that it is the owner of the NFT, and second, that the NFT has not expired. These restrictions safeguard that the address has permissions to retrieve the data recorded on the smart contract for each token (see also Figure 10). Therefore, access to the data product constructed to include all information produced in the Limassol farm is now granted to the owner of the corresponding address. If any other address besides the owner of the NFT attempts to retrieve that data, it is automatically blocked by the smart contract and it is not allowed to enter the token-gated website (see Scenario 3).

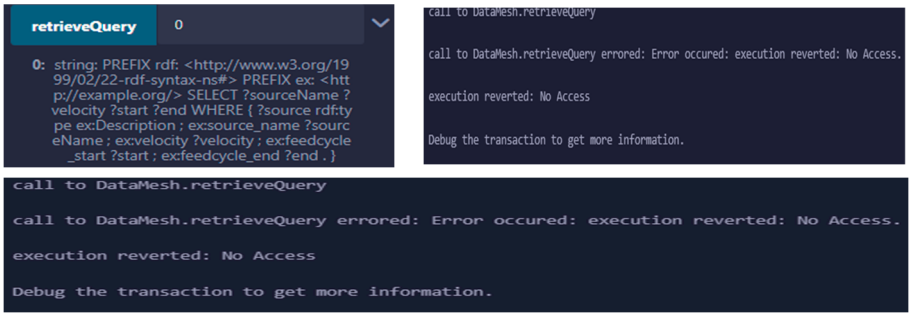

Figure 10. Query results directly from on-chain data and portal.

5.2.3. Scenario 3 – Applying Transfer Restrictions

This scenario demonstrates the transfer restrictions that are set by the administrator when a token is minted. As described in Scenario 1, token#0 was minted as a non-transferable token, while token#1 was minted with transferrable properties. As outlined in Figure 11, when the owner attempts to transfer token#0 to a different address it is blocked by the smart contract as this is not a valid action due to transfer restrictions. Subsequently, when the owner of token#1 tries to transfer the token, this is carried out successfully as token#1 has the appropriate transfer rights and hence the permissions to do so. Here the owner of token#1 transfers the token to address *0xC70bc32E46378B5a01c713d6dB18 042Acd8F0200*. Upon confirmation of the transaction on the Blockchain network, the previous owner of the token loses access to it as now the access rights are transferred to the new owner. Therefore, access to the data products is secured via Blockchain and single control of ownership is guaranteed by the NFT.

 doi:10.20944/preprints202403.0265.v1

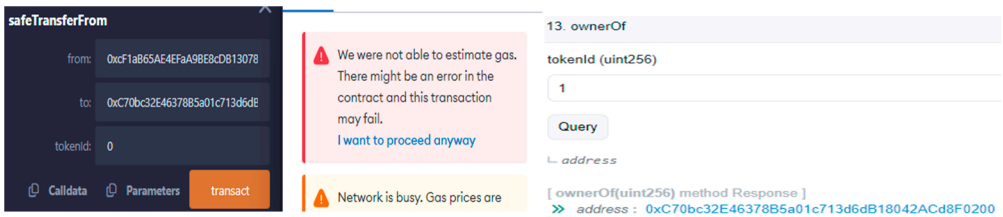

Figure 11. Result for token#0 and token#1.

6. Conclusions

This paper introduced an innovative framework for securing access and ownership in Data Meshes based on Blockchain and NFTs. The framework is applied on a Data Lake storage architecture, which may host Big Data at any scale, frequency and format, and utilizes Semantic Data Blueprints for dynamically constructing data products in Data Meshes. These products are designed to meet user demands and ensure that stakeholders access specific areas of the Data Mesh as needed through the transfer of ownership. The integration of NFTs and Blockchain technology offers a novel approach to address ownership and governance concerns. A dedicated algorithm was developed for incorporating the ability to mint time-based NFTs, thus facilitating secure and transparent data ownership transfers.

The proposed framework was demonstrated using a real-world case study from the smart manufacturing area. Specifically, a Data Lake was built to host data produced at a poultry meat production factory by several sensors and automated systems during the breeding process followed in the farms. Specific portions of data were selected to construct data products in a custom Data Mesh which were then used as key elements for granting access and transferring ownership to authorized users via the execution of smart contracts and NFTs. The scenarios tested suggested successful behavior in terms of ease of use, transparency, and correctness. It should also be noted that users in the factory (workers and managers in breeding sites and production line) were able to follow easily the algorithmic approach of the proposed framework and apply its steps efficiently, appreciating and greatly appreciated the ability to share data.

Future research steps will focus on enhancing and automating parts of the framework by utilizing recommender systems driven by user preferences and/or history of transactions for creating data products. Specifically, the system will be modified to employ advanced algorithms for analyzing user behavior and preferences to generate access recommendations. This will allow data products to be constructed upfront thus speeding up the process for granting access and transferring ownership.

References

1. Gupta, S.; Kar, A.K.; Baabdullah, A.M.; Al-Khowaiter, W.A.A. Big Data with Cognitive Computing: A Review for the Future. *International Journal of Information Management* **2018**, 42, 78-89, 10.1016/J.IJINFOMGT.2018.06.005.
2. Blazquez, D.; Domenech, J. Big Data Sources and Methods for Social and Economic Analyses. *Technological Forecasting and Social Change* **2017**, 130, 99-113, 10.1016/J.TECHFORE.2017.07.027.
3. Al-Sai, Z.A.; Husin, M.H.; Syed-Mohamad, S.M.; Abdin, R.M.S.; Damer, N.A.; Abualigah, L.; Gandomi, A.H. Explore Big Data Analytics Applications and Opportunities: A Review. *Big Data and Cognitive Computing* **2022**, 6(4), 157, 10.3390/bdcc6040157.
4. Khan, N.; Alsaqer, M.; Shah, H.; Badsha, G.; Abbasi, A.A.; Salehian, S. *The 10 Vs, Issues and Challenges of Big Data*. In Proceedings of the 2018 international conference on big data and education, 2018. 10.1145/3206157.3206166.
5. Khine, P.P.; Wang, Z. A Review of Polyglot Persistence in the Big Data World. *Information* **2019**, 10(4), 141., 10.3390/INFO10040141.
6. Shahid, A.; Nguyen, T.-A.N.; Kechadi, M.-T. Big Data Warehouse for Healthcare-sensitive Data Applications. **2021**, *Sensors*, 21(7), 2353, 10.3390/S21072353.
7. Driessen, S.; den Heuvel, W.J.V.; Monsieur, G. Promote: *A Data Product Model Template for Data Meshes.* In International Conference on Conceptual Modeling 2023, 125-142. Cham: Springer Nature Switzerland.
8. Kunigk, J.; Buss, I.; Wilkinson, P.; George, L., Architecting Modern Data Platforms: *A Guide to Enterprise Hadoop at Scale, O'Reilly Media*, 2018.

9. Derakhshannia, M.; Gervet, C.; Hajj-Hassan, H.; Laurent, A.; Martin, A. Data Lake Governance: Towards a Systemic and Natural Ecosystem Analogy. *Future internet*, 12(8), **2020**, 10.3390/FI12080126.
10. Pingos, M.; Andreou, A. *A Data Lake Metadata Enrichment Mechanism via Semantic Blueprints.* In 17th International Conference on Evaluation of Novel Approaches to Software Engineering, 2022, pp. 186-196, doi:10.5220/0011080400003176.
11. Pingos, M.; Andreou, A.S. Exploiting Metadata Semantics in Data Lakes Using Blueprints. *In International Conference on Evaluation of Novel Approaches to Software Engineering*. Publisher: Springer Nature Switzerland, 2022, pp. 220-242. doi:10.1007/978-3-031-36597-3_11.
12. Pingos, M.; Christodoulou, P. and Andreou, A. *DLMetaChain: An IoT Data Lake Architecture Based on the Blockchain*. In 13th International Conference on Information, Intelligence, Systems & Applications (IISA), July 2022, pp. 1-8. IEEE. doi:10.1109/IISA56318.2022.9904404.
13. Beheshti, A.; Benatallah, B.; Nouri, R. and Tabebordbar, A. CoreKG: a knowledge lake service. *Proceedings of the VLDB Endowment* **2022**, 11(12), pp.1942-1945.
14. Derakhshannia, M., Gervet, C., Hajj-Hassan, H., Laurent, A. and Martin, A. Data Lake governance: Towards a systemic and natural ecosystem analogy. *Future internet*, **2020** 12(8), p.126. doi:10.3390/FI12080126.
15. Dehghani, Z. How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh Available online: https://martinfowler.com/articles/data-monolith-to-mesh.html (accessed on 10 January 2024).
16. Dehghani, Z. Data Mesh Principles and Logical Architecture Available online: https://martinfowler.com/articles/data-mesh-principles.html (accessed on 10 January 2024).
17. Viriyasitavat, W.; Da Xu, L., Bi, Z. and Hoonsopon, D. Blockchain technology for applications in internet of things—mapping from system design perspective. *IEEE Internet of Things Journal*, **2019**, 6(5), pp.8155-8168. doi:10.1109/JIOT.2019.2925825.
18. Alam, T. Blockchain-based Internet of Things: Review, Current Trends, Applications, and Future Challenges. *Computers* **2022**, 12. doi:10.3390/computers12010006.
19. Di Angelo, M. and Salzer. *Tokens, types, and standards: identification and utilization in Ethereum.* In 2020 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS), August 2020, pp. 1-10. IEEE.
20. Yildiz, H.; Küpper, A.; Thatmann, D.; Göndör, S. and Herbke, P. Towards Interoperable Self-sovereign Identities. *IEEE Access* **2023**.
21. Rehman, W.; e Zainab, H.; Imran, J. and Bawany, N.Z. *NFTs: Applications and challenges.* In 2021 22nd International Arab Conference on Information Technology (ACIT), Dec 2021, pp. 1-7. IEEE.
22. Phuc, N.T., Khanh, H.V., Khoa, T.D., Khiem, H.G., Huong, H.L., Ngan, N.T., Triet, N.M., Kha, N.H., Anh, N.T., Bang, L.K. and Hieu, D.M. An Enhanced CoD System Leveraging Blockchain, Smart Contracts, and NFTs: A New Approach for Trustless Transactions. *International Journal of Advanced Computer Science and Applications*, **2023,** 14(10).
23. Shae, Z.Y.; Tsai, J.J. *On the Design of Medical Data Ecosystem for Improving Healthcare Research and Commercial Incentive.* In Proceedings of the 2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI), December 19, 202. IEEE.
24. Dolhopolov, A., Castelltort, A. and Laurent, A. *Implementing a Blockchain-Powered Metadata Catalog in Data Mesh Architecture*. In International Congress on Blockchain and Applications, July 2023, pp. 348-360. Cham: Springer Nature Switzerland.