

Transforming Data Lakes to Data Meshes Using Semantic Data Blueprints

Michalis Pingos^a, Athos Mina^b and Andreas S. Andreou^c

Department of Computer Engineering and Informatics, Cyprus University of Technology, Limassol, Cyprus

Keywords: Big Data, Data Lakes, Data Meshes, Data Products, Data Blueprints, Metadata Semantic Enrichment.

Abstract: In the continuously evolving and growing landscape of Big Data, a key challenge lies in the transformation of a Data Lake into a Data Mesh structure. Unveiling a transformative approach through semantic data blueprints enables organizations to align with changing business needs swiftly and effortlessly. This paper delves into the intricacies of detecting and shaping Data Domains and Data Products within Data Lakes and proposes a standardized methodology that combines the principles of Data Blueprints with Data Meshes. Essentially, this work introduces an innovative standardization framework dedicated to generating Data Products through a mechanism of semantic enrichment of data residing in Data Lakes. This mechanism not only enables the creation readiness and business alignment of Data Domains, but also facilitates the extraction of actionable insights from software products and processes. The proposed approach is qualitatively assessed using a set of functional attributes and is compared against established data structures within storage architectures yielding very promising results.

1 INTRODUCTION

In today's data-driven world, Big Data pervades every facet of our digital existence, while it is omnipresent and indispensable for producing insights that shape our world. It is the ubiquitous force driving innovation, analytics, and informed decision-making across diverse domains (Awan et al., 2021).


In essence, Big Data refers to extremely large and complex datasets that exceed the capabilities of traditional data processing methods and tools. Big Data originates from heterogeneous data sources with atypical patterns, which produce various kinds of structured, semi-structured, and unstructured data in high frequencies (Blazquez & Domenech, 2018). Big Data is often compared to gold as it offers the potential to yield valuable insights into various aspects of our daily lives. Through effective collection and analysis, it enables us not only to gain understanding but also to forecast future occurrences using predictive and prescriptive analytics.


The relevant scientific area has gained more attention as a result of revolutionizing technologies,


such as Internet of Things (IoT), which produce large amounts of data (Mehboob et al., 2022) and are applied in many areas such as Smart Healthcare, Smart Cities, Smart Grid, Smart Manufacturing etc.

As the concept of Big Data has evolved so rapidly, there has been some confusion regarding how it should be explained; this has led to a divergence in terminology between “what Big Data is” and “what Big Data does”. This evolving landscape underscores the challenges associated with comprehensively defining and understanding the multifaceted roles and functionalities of Big Data (Machado et al., 2022).

In the contemporary landscape of Big Data, the exponential growth in volume, variety, and complexity of data has necessitated the evolution of storage architectures to effectively manage and harness this wealth of information. Traditional storage solutions are often equipped to handle the diverse nature of modern data, which includes unstructured and semi-structured formats alongside conventional structured data. To address these challenges, innovative storage paradigms such as Data Lakes (DLs), Data Meshes (DMs) and Data

^a  <https://orcid.org/0000-0001-6293-6478>

^b  <https://orcid.org/0009-0007-6869-6090>

^c  <https://orcid.org/0000-0001-7104-2097>

Markets (DMRs) have emerged as indispensable components of the data infrastructure.

DLs act as expansive repositories capable of accommodating vast amounts of raw, unprocessed data in its native form. Meanwhile, DMs enable organizations to distribute and decentralize data processing, promoting scalability and flexibility. A DMR is an organized and structured platform or ecosystem where data is treated as a tradable commodity, enabling the buying and selling of datasets, information, or insights. These architectures empower businesses to derive insights from a broader spectrum of data types, fostering a more holistic and dynamic approach to data storage and analysis in the era of Big Data.

As previously mentioned, in the 2010s, DL architectures were introduced as structures well-suited for handling Big Data and guiding organizations towards a data-driven approach. Current research indicates a shift towards decentralized data exchange architectures, like DMRs and DMs (Driessen et al., 2021). Specifically, DMs aim to overcome certain limitations associated with monolithic data platforms like DLs (Dehghani, 2019). The development of effective data products imposes demands on metadata templates, which are currently not adequately addressed by existing methodologies.

The present paper deals with transforming a DL into a DM enjoying the benefits of rapidly storing high frequency data (DL) and constructing on-demand portions of information in the form of data products (DM). The proposed approach builds on the notion of data blueprints that aim at semantically annotating data before storing it in the DL. This metadata semantic enrichment guides the process for locating, retrieving and ultimately constructing data products easily and quickly according to user needs. The approach is demonstrated using two case studies. The first concerns real-world manufacturing data collected by a prominent local industrial entity in Cyprus and the second uses data obtained also from Europeana Digital Heritage Library (EDGL) and concerns cultural artifacts published by Europeana and accessed by the public. The data from the two case studies are stored in a dedicated DL utilizing the proposed semantic metadata enrichment mechanism. Subsequently, the DMs are produced centred around diverse data products. Performance is then evaluated by varying the complexity of the data products constructed based on the granularity of information sought and the number of data sources involved.

The remainder of the paper is structured as follows: Section 2 presents the technical background of the paper and section 3 discusses briefly the related

work and the areas of DLs and DMs. Section 4 presents the framework for transforming DLs into DMs and discusses the contribution of semantic data blueprints. Experimentation conducted to assess performance is showcased in section 5, wherein a set of experiments is designed and executed using real-world data acquired from PARG and EDGL. Finally, section 6 concludes the paper and outlines potential future research steps.

2 TECHNICAL BACKGROUND

A DL is one of the debatable ideas that emerged during the Big Data era. DLs were proposed in 2010 by James Dixon, Chief Technology Officer of Pentaho, as architectures suitable for dealing with Big Data and for assisting organisations towards adopting data-driven approaches. DL is a relatively recent concept with groundbreaking ideas that has emerged in the past decade, bringing forth various challenges and obstacles to widespread adoption (Khine & Wang, 2018).

A DL serves as a centralized repository capable of storing structured, semi-structured and unstructured data at any scale. AWS (2022) defines a DL as a storage system where data can be stored in its raw form without the need for prior structuring. This enables the execution of various analytics, ranging from dashboards and visualizations to Big Data processing, real-time analytics, and machine learning, facilitating informed decision-making. The architecture of DLs extends to include the storage of both relational and non-relational data, seamlessly combining them with traditional Data Warehouses (DWs) for comprehensive data management.

The current literature shows a growing trend towards decentralized data exchange systems, exemplified by concepts such as DM. Coined by Zhamak Dehghani in 2019 during her tenure as a principal consultant at ThoughtWorks, the term "Data Mesh" encapsulates a paradigm shift towards more distributed and collaborative approaches to managing and sharing data.

DM is a revolutionary concept in data architecture that aims to tackle the challenges presented by centralized data systems. Essentially, a DM advocates for a decentralized approach wherein data ownership and governance are distributed among various autonomous domains within an organization which are offered through APIs. This innovative structure promotes the formation of cross-functional teams with domain-specific expertise, fostering accountability and

a sense of duty towards their respective areas (Wieder & Nolte, 2022).

In this paradigm shift, each team becomes self-sufficient, responsible for managing their own data infrastructure, storage solutions and processing capabilities. By doing so, scalability and agility in handling vast amounts of information are fostered.

Furthermore, it is important to note that while DMs is more about organizational and conceptual principles for data management, DLs refer specifically to the technology and infrastructure for storing large volumes of raw data. These concepts are not mutually exclusive, and, in practice, they may coexist as organizations can implement a DM framework while utilizing a DL as one component of their technical infrastructure for data storage and processing.

3 RELATED WORK

The exploration of transforming DLs to DMs appears in the international literature to be in its early phases, suggesting that this paradigm shift in data architecture is not yet a well-established or widely adopted concept. This novelty is evident in the limited number of publications and scholarly works addressing the topic, many of which are very recent, indicating a surge in interest. The challenge lies in the intricate nature of this transformation, as transitioning from traditional monolithic DLs to a distributed DM involves complex changes in data ownership, architecture, and organizational structures.

The fundamental concepts and principles behind the DM paradigm signal a significant shift in data architectures (Machado et al., 2022). That paper delved into the core principles of DM, such as decentralized data ownership, treating data as a product, and advocated for a federated and domain-oriented approach to handling data at scale within organizations. The paper also provided insights into the conceptual framework and guiding principles for implementing DMs as an innovative approach to managing and leveraging data assets.

Furthermore, Driessen et al., (2023) introduced a data product model template named ProMoTe, designed specifically for DM architectures. The authors proposed a structured framework to guide the implementation of data products within the context of DM addressing key aspects such as ownership, discoverability, and scalability. The paper contributes to the growing body of literature on DM by providing a practical tool or model for organizations looking to adopt and implement the principles of DM in their data architectures.

The integration of DM and microservices principles to form a cohesive and unified logical architecture is explored by Morais et al. (2023). The authors highlighted how the decentralized, domain-oriented principles of DMs can be harmonized with the modular and scalable nature of microservices. The paper proposed a unified model that leverages the strengths of both DM and microservices to create a comprehensive and adaptable solution for managing large-scale data ecosystems.

Dehghani (2019) argues that traditional monolithic DLs often face challenges related to scalability, agility and ownership, and proposes a distributed across autonomous, cross-functional teams.

The architecture proposed in this paper for transforming DLs to DMs adopts the Semantic Data Blueprints concept reported in (Pingos and Andreou, 2022). The authors of that paper presented a mechanism for enriching metadata in a DL through the use of semantic blueprints as an extension of manufacturing blueprints presented earlier (Papazoglou and Elgammal, 2018). The authors actually proposed a method to enhance the metadata within a DL environment, leveraging semantic blueprints as a guiding framework of describing data sources before they become part of a DL.

The mechanism introduced in that work involved incorporating semantic structures and utilizing both the theory of Triples (subject-predicate-object) and the Resource Description Framework (RDF) to improve the organization, mapping, and retrieval of data stored in the DL. The paper essentially provided insights into how semantic blueprints can be utilized in conjunction with the 5Vs characteristics of Big Data to improve the effectiveness and metadata quality within DLs, addressing challenges associated with managing and extracting meaningful information from large and diverse datasets.

The aforementioned framework was extended in (Pingos and Andreou, 2022) which explored the development of a metadata framework for process mining within the context of smart manufacturing utilizing DLs in a smart factory. The authors proposed a systematic approach for paradigm where data is treated as a product and is organizing and enhancing metadata to support process mining activities in smart manufacturing environments by introducing process blueprints. That work also contributed to the design and implementation of a metadata framework using semantic blueprints tailored for smart manufacturing DLs, aiming to improve the efficiency and effectiveness of process mining activities in this context.

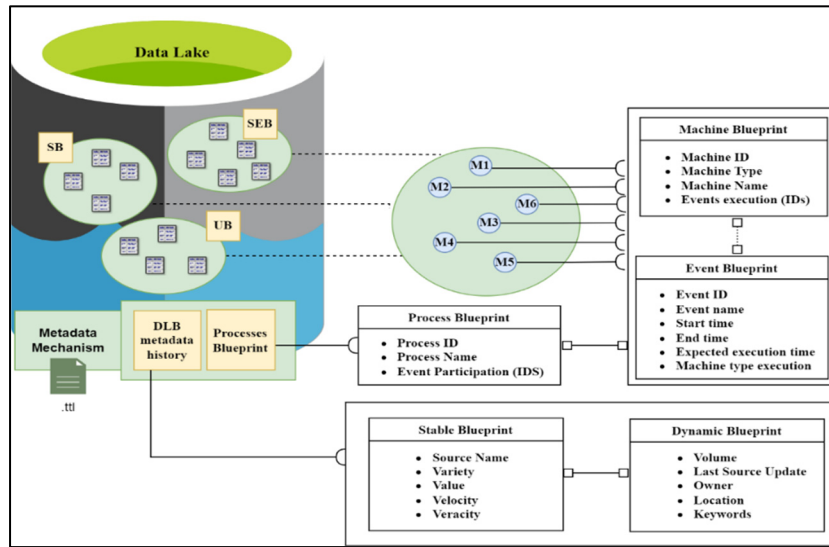


Figure 1: Overview of Semantic Data Blueprints framework.

The envisioned architecture for the DL comprised various data ponds, each dedicated to hosting or referencing a distinct data type based on its designated design. Each pond featured an exclusive data processing and storage system tailored to the nature of the data it accommodated. The proposed approach underwent a comparative analysis with existing metadata systems, evaluating its efficacy based on a set of functional attributes indicating that it constitutes a promising and viable strategy. Figure 1 summarizes the aforementioned frameworks and provides an overview of the concept of Semantic Data Blueprints and all extensions or enhancements performed on them, which were adopted by our paper in order to transform DLs to DMs and produce data products and levels according business needs.

Further to the above, a comprehensive state of the art of the different approaches to DLs design was provided by Sawadogo and Darmont (2021). They particularly focused on DL architectures and metadata management, which are key issues in successful DLs. The authors delved into the intricacies of DL design, storage mechanisms, and processing capabilities, emphasizing the challenges posed by the vast and diverse datasets they store. The article stated also the importance of effective metadata management for enhancing data governance, ensuring quality, and supporting analytics.

Majchrzak et al. (2023) explored the practical implementation of DMs, focusing on key drivers for transformation decisions. The latter emphasized the integration of DLs and DWs in diverse formats within

the context of data meshes. Additionally, it addressed the relevance of related work in the field.

In addition, Holfelder et al. (2023) emphasized the ingestion of data into a DL, followed by processing to ensure compatibility. The authors suggested that this approach represents a shift in data architecture, sparked by the evolution of DMs not only by complementing traditional data warehouse architectures but also by bringing about a transformative impact on the overall data landscape. That work contributes valuable insights into the use of sovereign Cloud technologies for building scalable data spaces, providing a novel perspective on data processing and storage within the evolving context of data architecture.

Finally, Ashraf et al. (2023) explored the application of key lessons derived from microservices principles to facilitate the adoption of DMs. The authors highlighted the shift away from the conventional practice of centralizing data consumption, storage, transformation, and output.

The short literature overview provided in this section reveals that the challenge of transforming a DL into a DM is yet to be tackled and that there is ample room for approaches to address this issue in a standardized manner so that the benefits of the DLs are preserved and the advantages of producing data products within a DM are exploited. This is exactly what this paper does; it provides an efficient and flexible approach to creating DMs out of semantically annotated DL data.

4 METHODOLOGY

A novel standardization framework is introduced in this work aimed at transforming a DL into a DM through the utilization of Semantic Data Blueprints as presented in Figure 2. This framework leverages standardized data descriptions in the form of blueprints, employing a domain-driven approach to generate data products. To demonstrate the effectiveness of this approach, two case-studies are used, one from the domain of smart manufacturing and the other of cultural heritage.

A typical DL is employed here which is further enhanced with metadata mechanism which essentially describes the sources that produce the data residing in the DL. This metadata plays a crucial role in providing the transformation of the DL to DM and is constructed using .ttl files, the latter referring to Turtle, a widely used serialization format for RDF data. The .ttl files, through the use of the Turtle syntax, enable the creation of structured and semantically rich metadata within the DL. This enhances comprehensibility and accessibility of the data by offering a standardized and machine-readable representation of the metadata, facilitating efficient data management and utilization within the DL environment.

The ability to create Data Products and Data Domains while transforming a DL to DM is based on a dedicated form of blueprint as presented in the DL metadata description examples in GitHub link (<https://github.com/mfpingos/ENASE2024>), which provides examples of source descriptions within a .ttl file that correspond to data produced from PARG factory and EDGL.

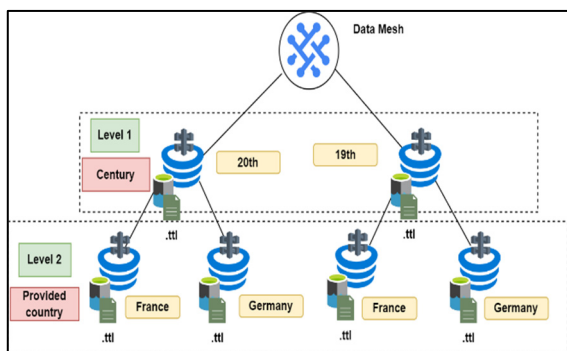


Figure 2: Creation of DM Data Domains according to data owner needs.

PARG is a prominent local industrial entity, recognized as one of the key players and leading authorities in the domain of poultry farming and the trade of poultry meat in Cyprus. The company provides an extensive range of top-

notch products designed to cater to the contemporary consumer's preferences for convenient cooking and healthy dietary choices.

EDGL is a well-known cultural heritage website that provides access to cultural heritage materials such as libraries, museums, archives, and other cultural institutions across the continent. The European Commission launched the Europeana platform in 2008 to increase public access to Europe's cultural heritage. Millions of digital artifacts, including books, artworks, pictures, manuscripts, maps, sound recordings, and archive documents, are available on this platform. The Europeana website offers users the ability to search for and view cultural goods having free access to them. Virtual exhibitions, educational resources, and APIs for developers are just a few of the additional tools and services that Europeana provides to assist users in exploring and interacting with the materials.

The manufacturing data and business processes are confidential, and the digital heritage items are protected by intellectual property rights. Therefore, the present work made every effort to preserve data confidentiality where appropriate. In the case of the PARG factory, data underwent masking or downgrading to ensure anonymity and business confidentiality during demonstrations and when sharing descriptions. In the case of EDGL, synthetic data is generated using the existing metadata descriptions that are already available on the website.

Despite applying the above measures, the provided case studies are able to sufficiently illustrate the fundamental principles of the proposed framework, demonstrating its applicability and usefulness as described above. It should also be noted that the cases were selected so as to demonstrate the wide applicability of the framework irrespective of the application domain or data involved.

Table 1: Experimentation DMs levels for the PARG.

	PARG
Level 2	Location, Variety
Level 3	Location, Variety, Velocity
Level 4	Location, Variety, Velocity, Flock size
Level 5	Location, Variety, Velocity, Flock size, Year
Level 6	Location, Variety, Velocity, Flock size, Year, Sensors Accuracy

As mentioned above, a DL was constructed using the metadata mechanism for semantic annotation of the sources. In order to transform the DL to DM as presented in Figure 2, a dedicated middleware was

developed which has been installed on a server and is being fed with user/owner preferences (uploaded on GitHub link given earlier). In essence, the owner defines the data products, and, hence, the levels of the DM according to business needs.

Table 2: Experimentation DMs levels for the EDGL.

	EDGL
Level 2	variety, theme
Level 3	variety, language, theme
Level 4	variety, language, format, theme
Level 5	variety, language, format, rights, theme
Level 6	variety, language, format, type_of_object, rights, theme

For example, as can be observed in Europeana’s website, each registered digital item is characterized with a specific metadata structure (examples uploaded in GitHub link given earlier). In order to demonstrate the proposed framework, we selected the following eight significant metadata characteristics: Century, Providing Institution, Type of object, Subject, Identifier, Places, Format, Providing Country.

Let us now assume that using the aforementioned description the owner of the data wishes to create two levels for the DM and set Century and Providing country as the preferred characteristics. The example of the DL .ttl file consists of items of 19th century and items for 20th and provided by Germany and France. The metadata description is pushed to middleware as a result according the data owner needs the DM created as presented in Figure 2. In essence every part of the DM is a DL that consist of metadata according the defined levels of the user. The selected level attributes are sourced by the cultural heritage metadata characteristics of the DL. These are treated as the components of the Data Mesh architecture providing the ability to create Domains according to selected attributes expressed via the data blueprint mechanism introduced (Pingos and Andreou, 2022)

The next section demonstrates the applicability and effectiveness of the proposed framework, which is also evaluated by converting the initial DL into a DM creating various data products (levels) using the PARG and EDGL metadata description. Note that the metadata mechanism describes the sources characteristics defining also the location of each source in the DL. Finally, the framework is assessed by executing and comparing the performance of queries based on the DM level and using the metadata mechanism directly on the DL.

5 EXPERIMENTAL VALIDATIONS

5.1 Design of Experiments

The experiments conducted had dual objectives. Firstly, they sought to assess the capability of the proposed approach in generating DMs and refined data products/levels through the utilization of Semantic Data Blueprints. Secondly, the experiments aimed to evaluate the performance and effectiveness of the approach in terms of granularity. To fulfil these objectives, a series of experiments were carried out, and this section provides an explanation of the rationale behind their design.

Data from two different application areas, smart manufacturing (PARG) and digital heritage (EDGL) were utilized for the execution of the experiments. As a starting point, a DL metadata mechanism was built for each area (uploaded also in Github). The DL metadata was described with a .ttl file which contains the characteristics for each source that stored data in the DL.

As mentioned in the previous sections, Python scripts automatically created the .ttl files while also masked sensitive data. The growth in the number of sources directly impacts (increases proportionally to) the size of the respective .ttl file, a crucial element parsed to extract sources that match a query. As an illustration, in the PARG case a .ttl file describing 100 sources resulted in a size of 0.077 MB, 1000 sources produced 0.769 MB, 10000 sources yielded 7.5 MB, and 100000 sources led to a file size of 75.9 MB. In the case of EDGL, 100 sources in a .ttl file resulted in a size of 0.103 MB, 1000 sources amounted to 1 MB, 10000 sources equated to 10.1 MB, and 100000 sources reached 101.7 MB.

However, it must be noted that despite the similar number of sources in the DL example for each application area, there is a variance of the respective file sizes. This difference arises because the EDGL sources’ description includes more attributes, specifically, EDGL sources are described with 20 attributes, whereas PARG sources are described with 15 attributes (also indicated when following the GitHub link). The size of the initial DL metadata characteristics and the number of attributes represent another aspect explored in the experiments.

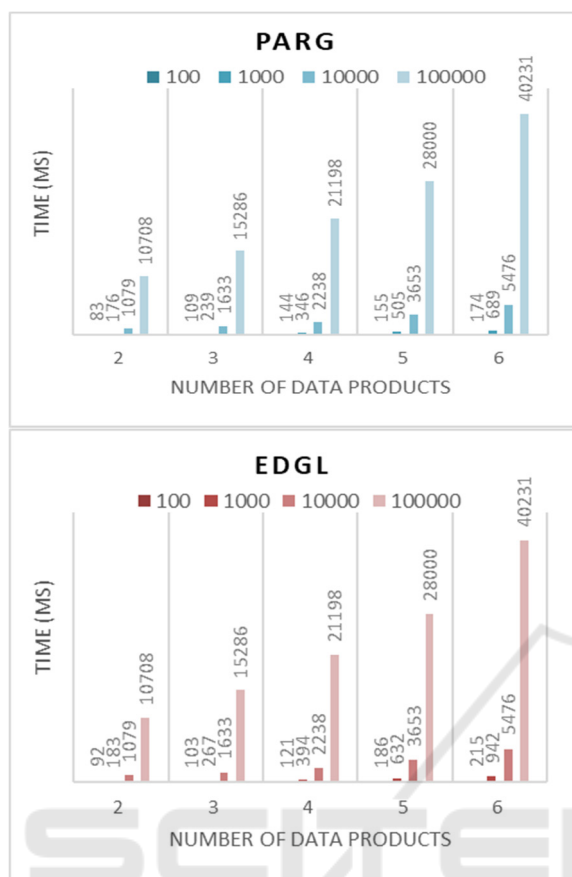


Figure 3: Time performance for constructing data products with different numbers of data sources for the two use-cases.

The experiments were conducted on a server computer comprising three Virtual Machines. The CPU configuration consisted of 4 dedicated cores, while the underlying server hosting these machines featured a total of 48 cores. The memory size allocated was 8192MB, and the hard disk capacity stood at 80GB. The software stack employed for the experiments included Hadoop (version 3.3.6) for distributed computing, Python (version 2.7.5) for scripting purposes, data generation based on raw real-world data from PARG and EDGL, and the creation of data products at the DM level. Additionally, Apache Jena was utilized for SPARQL query processing.

Two queries were constructed and executed using all DL descriptions sizes and all DM levels produced. The first query (Query1.sparql) is executed on PARG and selects values for variables flockid, source_name, and source_path, where the RDF triples match a set of conditions. The conditions include the accuracy of sensors being “Medium”, the location “Limassol”, the data variety “Structured”, the velocity “Hourly”,

the flock size being “Low”, and the year “2020”. These criteria indicate a focus on data related to a specific context, pertaining to sensor information associated with a flock, with additional constraints on the geographical location, data characteristics, temporal aspects, and other specific attributes.

The second SPARQL query (Query2.sparql) executed on EDGL metadata is formulated also to extract specific information from the .ttl file based on specified conditions. In essence, the query selects values for variables: providing institution, source_path and provider collection name. The conditions set for retrieval include criteria such language being “De” (German), data variety “Unstructured”, format “audio/mp3”, type of object “3D”, rights associated with the “Creative Commons” license and a thematic association with “Manuscript”.

The queries were structured to retrieve and display relevant data that meets the aforementioned defined criteria, both with the same complexity in order to be comparable. Note that the queries are executed to the .ttl file of the last (maximum) data product level provided by the corresponding DM structure.

5.2 Experimental Results

Figure 3 illustrate the time required for constructing the DM levels in each application domain as presented in Table 1 and Table 2. DLs with varying metadata size and number of data sources were transformed into DMs with diverse granularity levels and data products. As indicated in the case studies, the transformation time escalates in alignment with both the number of sources and the attributes characterizing those sources.

The creation time for DMs with 2 levels is minimal, and this time steadily increases as more data products (granularity levels) are generated based on data owner requirements, as expected. The construction time for DMs with the maximum level (6 data products) is significantly higher in the two examples compared to lower levels, exhibiting an average increase between 3 and 10 times as the number of data sources is increased for the same number of DM levels created in both case-studies. It is noteworthy that the maximum construction time for DMs is less than 0.6 minutes for PARG DL metadata and less than 0.7 minutes for EDGL metadata. This can be regarded as a quite satisfactory performance, especially considering the extreme conditions tested with values reaching 100,000 for the sources and 6 for the granularity levels that are, in practice, very rare to encounter. This also indicates that the number of

attributes describing the sources affects the construction time, as expected, because it increases the .ttl file size.

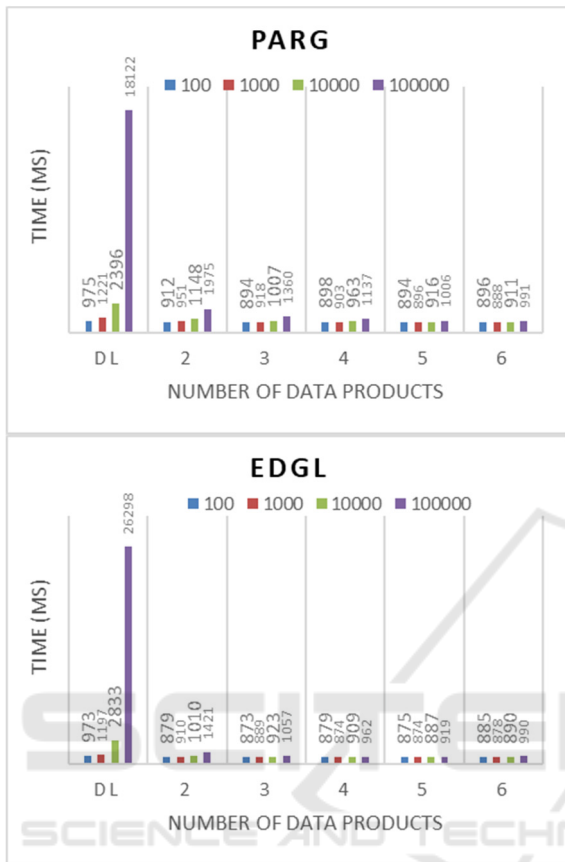


Figure 4: Time performance for executing queries on DMs with varying number of data products and data sources for the two use-cases.

The two benchmark SPARQL queries were executed 100 times each using PARG’s and EDGL’s metadata and different DL and DM structures produced by varying the number of sources, and hence the metadata in the .ttl files, to facilitate a comprehensive comparative analysis. Figure 4 present the query execution times in milliseconds, accompanied by the corresponding number of sources. To ensure a standardized comparison, the queries were configured to yield an identical number of sources at each level.

Notably, the observed trend reveals a direct correlation between query execution time and the aggregate number of sources returned. Specifically, as the granularity increases (i.e. the number of data products), there is a discernible decrease in query execution time. This observation highlights a significant advantage inherent in employing a DM

structure utilizing Semantic Data Blueprints, that is, the capacity to confine information within designated data product levels thereby facilitating immediate and efficient data retrieval.

Finally, it is evident that maximizing the granularity, if needed, in constructing DMs proves beneficial. This becomes particularly apparent when comparing the execution time of a query on a DL with 100,000 sources, as an extreme scenario, against a DM Level 6 with the same number of sources which both return 59 sources satisfying the query for PARG and 8 sources for EDGL. The query execution time is observed to be 18 times faster in the latter case for PARG and 26,5 times faster for the EDGL.

6 CONCLUSIONS

This paper explored the conversion of a Data Lake (DL) into a Data Mesh (DM), leveraging the advantages of efficiently storing high-frequency data (DL) and constructing specific information segments as data products. The proposed approach was based on the concept of data blueprints, which involve semantically annotating data before storing it in the DL. This semantic enrichment of metadata guides the process of locating, retrieving, and swiftly constructing data products based on user requirements. The approach was exemplified through two case studies.

The first employed real-world manufacturing data from the Paradisiotis Group of Companies (PARG), a prominent local industrial entity in Cyprus focusing on poultry farming and poultry meat product production and trading. The second case study utilized data from the Europeana Digital Heritage Library (EDGL), specifically cultural artifacts published by Europeana and accessed by the public.

The data from both case studies were stored in a dedicated DL using the proposed semantic metadata enrichment mechanism. Subsequently, DMs were generated, centered around various data products defined by the user. The performance was then assessed by varying the complexity of the constructed data products based on the granularity of information sought and the number of data sources involved.

The target of the conducted experiments was twofold: Firstly, they aimed to evaluate the capability of the proposed approach in generating DMs and refined data products/levels through the application of Semantic Data Blueprints. Secondly, the experiments sought to assess the performance and effectiveness of this approach concerning granularity.

The results obtained were quite satisfactory indicating that transforming a DL to DM is fully supported under the proposed semantic enrichment mechanism with limited time requirements, as well as consistent behaviour over varying number of sources residing in the DL and complexity of the queries executed to retrieve these sources.

Future work will concentrate on decentralization of data ownership and access of the DM created using the transformation approach proposed here by using Blockchain and NFT technology. DM, as presented also in this paper, is a methodology for structuring and managing data by considering it as one or more products and emphasizes the decentralization of data ownership and access. The latter has emerged as a topic posing numerous challenges concerning data ownership, governance, security, monitoring, and observability. To tackle these challenges, this framework will be extended to facilitate on-the-fly generation of DM and Data Products in response to user requests through visual queries, guaranteeing that stakeholders can access particular segments of the DM as dictated by their privileges, paving the way for the realization of Data Markets (DMRs).

REFERENCES

- Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, 130, 99-113.
- Awan, U., Shamim, S., Khan, Z., Zia, N., Shariq, S., & Khan, M. (2021). Big data analytics capability and decision-making: The role of data-driven insight on circular economy performance. *Technological Forecasting and Social Change*.
- Mehboob, T., Ahmed, I. A., & Afzal, A. (2022). Big Data Issues, Challenges and Techniques: A Survey. *Pakistan Journal of Engineering and Technology*, 5(2), 216-220.
- Machado, I. A., Costa, C., & Santos, M. Y. (2022). Data mesh: concepts and principles of a paradigm shift in data architectures. *Procedia Computer Science*, 196, 263-271.
- Driessen, S. W., Monsieur, G., & Van Den Heuvel, W. J. (2022). Data market design: a systematic literature review. *IEEE access*, 10, 33123-33153.
- Dehghani, Z. (2019). How to move beyond a monolithic data lake to a distributed data mesh. *Tech. Rep.*
- Khine, P. P., & Wang, Z. S. (2018). Data lake: a new ideology in big data era. In *ITM web of conferences* Vol. 17, p. 03025. EDP Sciences.
- Amazon Web Services. (2022). What is a data lake? Retrieved from: <https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/>
- Wieder, P., & Nolte, H. (2022). Toward data lakes as central building blocks for data management and analysis. *Frontiers in Big Data*, 5.
- Driessen, S., den Heuvel, W. J. V., & Monsieur, G. (2023). ProMoTe: A Data Product Model Template for Data Meshes. In *International Conference on Conceptual Modeling*, 125-142. Cham: Springer Nature Switzerland.
- Morais, F., Soares, N., Bessa, J., Vicente, J., Ribeiro, P., Machado, J., & Machado, R. J. (2023, September). Converging Data Mesh and Microservice Principles into a Unified Logical Architecture. In *International Conference on Intelligent Systems in Production Engineering and Maintenance*, 300-314. Cham: Springer Nature Switzerland.
- Pingos, M., & Andreou, A. S. (2022). A Data Lake Metadata Enrichment Mechanism via Semantic Blueprints. In *17th International Conference on Evaluation of Novel Approaches to Software Engineering ENASE*, 186-196.
- Papazoglou, M. P., & Elgammal, A. (2017). The manufacturing blueprint environment: Bringing intelligence into manufacturing. In *2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 750-759. IEEE.
- Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, 56, 97-120.
- Pingos, M., & Andreou, A. S. (2022). A Smart Manufacturing Data Lake Metadata Framework for Process Mining. *The Nineteenth International Conference on Software Engineering Advances ICSEA*, 11.
- Majchrzak, J., Balnojan, S., & Siwiak, M. (2023). *Data Mesh in Action*. Simon and Schuster.
- Holfelder, W., Mayer, A., & Baumgart, T. (2022). Sovereign Cloud Technologies for Scalable Data Spaces. *Designing Data Spaces*, 419.
- Pingos, M., & Andreou, A. S. (2022). Exploiting Metadata Semantics in Data Lakes Using Blueprints. In *International Conference on Evaluation of Novel Approaches to Software Engineering*, 220-242. Cham: Springer Nature Switzerland.