METADATA MANAGEMENT FRAMEWORK FOR BUSINESS INTELLIGENCE DRIVEN DATA LAKES

Snezhana Sulova[1],
Olga Marinova[2]

Abstract: Data lakes (DL) provide powerful capabilities for processing and utilizing large and diverse data, helping organizations adapt to the modern environment and extract maximum value from the information at their disposal. Effective data analysis provides actionable knowledge which is a competitive advantage for organizations. Metadata management in data lakes is a key element in ensuring their full functionality. At the same time, this is a dynamic and under-researched area that reflects the rapid development of information technology and the business needs for effective data management. The research is based on a thorough scientific analysis of existing publications on the chosen topic. For this purpose, up-to-date and relevant open access publications from Scopus and Web of Science that correspond to the keywords "data lake" and "metadata" are identified and are from the last 15 years. Based on a review of the existing literature, the main challenges in data lake metadata management are highlighted. The goal of the research is to summarize the existing models in the field of metadata management in data lakes and to propose a new conceptual framework that can serve as a useful guide for designing and implementing metadata management models in heterogeneous data warehouses, as well as implementation steps. The concept's adoption involves a detailed study of the data management model in a specific organization, a measurement of the level of effectiveness after the model's implementation, and the use of additional metrics to confirm its feasibility. These tasks are therefore the subject of future research. Another limitation of the proposed framework is that it does not address in depth the rules and standards related to ensuring data security, which would be of the highest priority especially in sectors such as finance, defence and healthcare. In addition, further research could also focus on future analysis of the level of satisfaction with the transformation of metadata management processes.

Key words: data lake, metadata, metadata management, data lake architecture, conceptual metadata framework.

JEL: C8.

DOI: https://doi.org/10.58861/tae.bm.2024.2.02

[1] Department of Informatics, University of Economics – Varna, e-mail: ssulova@ue-varna.bg, ORCID: 0000-0003-4889-0973

[2] Department of Informatics, University of Economics – Varna,e-mail: olga.marinova@ue-varna.bg, ORCID: 0009-0008-9026-7097

Introduction

Big data and its effective use have undoubtedly been a subject of established research and studies in recent years. Today, data has become a core business asset (Cristescu et al., 2023). Data in the Internet space is growing at an extraordinary pace, its heterogeneity is increasing, which, in turn, significantly complicates its extraction, unification and application for various analytical services in real time. It can be said that in order to effectively organize and manage the information coming daily from heterogeneous sources, the need for solutions for integrated access to all data in the organization is increasingly urgent. A common solution for storing and organizing large and varied types of data is the use of *data warehouses*.

Although *data warehouses* are still effectively used to store high-throughput structured data, storing semi-structured and unstructured data poses a significant challenge for them (Sawadogo & Darmont, 2021). The fact that the majority of data nowadays is unstructured (Bankov, 2018; Miloslavskaya & Tolstoy, 2016), proves that a new approach to dealing with the problems that accompany this huge and heterogeneous data is required.

In this regard, the relatively new concept of a “data lake” – an approach that focuses on the storage and management of big data, especially data which is unstructured, has gained popularity. A data lake can be thought of as a centralized repository where data from different formats and sources are stored without a strict pattern for recording them and with the idea of using them for future analysis (Couto et al., 2019; Khine & Wang, 2017). Two key characteristics of data lakes stand out in this definition: the heterogeneous nature of the data and the so-called “schema-on-read” approach, which means that the definition of the schema, the integration and transformation of the data is done as needed at the time of data access or “on demand” (Chihoub et al., 2020; Khine & Wang, 2017). This is in contrast to the traditional "schema-on-write" approach used in data warehouses, where the schema is defined before the data is integrated and the expected information is known in advance.

However, this research does not aim to consider data lakes as an alternative to data warehouses and highlight their differences, but rather explore the capabilities and key benefits of data lakes, and identify some practices that would make their use in the context of **business intelligence** more effective. The idea behind a data lake is to create a specialized repository that can collect different types of data without structuring it first (Sawadogo & Darmont, 2021; Derakhshannia et al., 2020). This allows

organizations to preserve their large volumes of data in their original format without losing information and valuable data.

Whenever data comes from heterogeneous sources with different models and formats, maintaining metadata is necessary to track the life cycle of that data. **Metadata** contains information about the original data, including the information schema, semantics and origin, as well as other relevant details.

The importance of metadata management to prevent the data lake from becoming a “data swamp” or useless data is emphasized by many authors (Diamantini, et al., 2018; Hai, et al., 2016). Some authors even define metadata management as the only possibility to guarantee an effective and efficient management of data source interoperability (Diamantini, et al., 2018). A **BI-driven data lake** refers to the concept where the data lake architecture and data management are primarily focused on the **specific needs of BI**. Building such a data lake aims to support and facilitate business analysis and decision making.

The global metadata management tools market size is estimated at US$ 6.68 billion in 2021 and is expected to grow at a compound annual growth rate (CAGR) of 20.8% during the period 2022 - 2030 (Metadata Management Tools Market Size, n.d). This trend shows that the field under consideration is extremely relevant and will continue to expand its scope and applicability in the future.

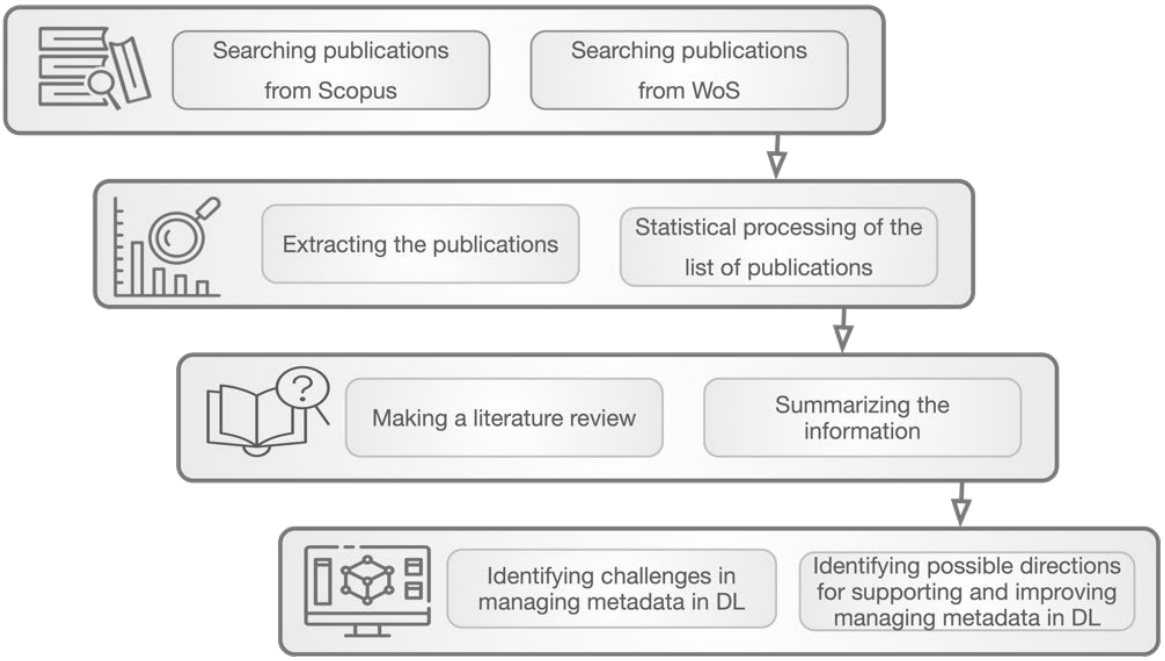
In this regard, the **main goal** is to summarize the existing models in the field of metadata management in data lakes and to create a conceptual framework that serves as a useful guide for designing and implementing an effective model for their management.

1. Research methodology

The chosen research methodology mainly applies theoretical research methods and a systematic literature review related to DL metadata management. In order to achieve the defined goal stated in section 1, the methodology presented in Figure 1 is followed by the authors.

The first stage of current research is the process of searching for relevant publications. Publications in the field of computer science, which are in the prestigious Scopus and Web of Science databases and respond to a search request using the keywords "data lake" and "metadata" are examined.

We have taken only the open access publications, or this is 72% of all found (47 publications from Scopus and 39 from Web of Science).



Source: own elaboration

Figure 1. Research methodology

The second stage is related to the retrieval of the publications found, and a list containing the name of the publication, year of publication, abstract and list of citations are created. MS Excel is used to process the lists – merging, removing duplicate publications and filtering only publications that have citations.

The next stages of this research are based on **theoretical analysis, synthesis** and **summary** of the main ideas of the publications found. After a comprehensive review, the most relevant publications for the study area are summarized in Table 1.

The studies reviewed in Table 1 can be divided into two main groups: publications that examine in detail the metadata management process in heterogeneous DL repositories (Hai et al., 2023; Nambiar & Mundra, 2022; Sawadogo & Darmont, 2021; Ravat & Zhao, 2019) and those, in which the emphasis is on offering an approach, a model for metadata storage and management (Cherradi & El Haddadi, 2023; Francia et al., 2021; Megdiche et al., 2021; Scholly et al., 2021; Armbrust et al., 2020; Eichler et al., 2021; Holom et al., 2020; Sawadogo et al., 2019a; Sawadogo et al., 2019b; Nogueira et. al., 2018;

Prabhune et al., 2018; Alserafi et al., 2016; Quix et al., 2016). Although the presented approaches are for DL, some of them are more suitable for structured data (Quix et al., 2016), and others are for specific data, for example, for extracting contextual information from files (Diamantini et al., 2018), and there are also more comprehensive models (Sawadogo et al., 2019b).

Table 1.
Major publications on metadata management in DL according to their scope

Reference	The DL concept and metadata	Metadata Management in DL	Types of metadata in DL	Model for Metadata Manage-ment in DL	Technologies for the implement-tation of Management in DL
Sawadogo & Darmont (2021)	x	x	x	-	-
Armbrust et al. (2020)	x	-	-	x	x
Ravat & Zhao (2019)	x	x	x	-	-
Sawadogo et al. (2019b)	x	x	x	x	-
Alserafi et al. (2016)	x	x	-	x	x
Sawadogo et al. (2019a)	x	-	-	x	x
Francia et al. (2021)	x	x	-	x	x
Eichler et al. (2021)	x	x	x	x	-
Nogueira et. al. (2018)	x	x	-	x	x
Nambiar & Mundra (2022)	x	x	-	-	x
Prabhune et al. (2018)	-	x	-	x	x
Scholly et al. (2021)	x	x	-	x	x
Quix et al. (2016)	x	-	-	x	x
Cherradi & El Haddadi (2023)	x	x	x	x	x
Hai et al. (2023)	x	x	x	-	-
Megdiche et al. (2021)	x	-	-	x	x
Hellerstein et al. (2017)	x	x	-	x	x
Holom et al. (2020)	-	x	-	x	x
Diamantini et al. (2018)	x	x	x	x	-
Skluzacek, et al. (2018)	-	-	-	x	x

Source: own elaboration

Based on the study and analysis of existing publications in the field under consideration, separate aspects of different theoretical concepts are combined; the key features of DL and the types of metadata in DL are defined; the challenges in their storage and management are identified (in Section 2), and a new conceptual framework for the maintenance and development of the metadata management process in the DL is proposed (in Section 3).

2. Metadata – types and challenges

Metadata is a key component in data lakes to ensure that information will continue to exist and be accessible in the long term. It has been identified as the key to understanding and manipulating data (Chen, 2022). They are particularly important for organizations, as they support the following areas: selection of appropriate information resources, organization of information, interoperability and integration, unique digital identification, archiving and data protection.

There are two main classifications of metadata in the literature – according to the functional purpose of the data (Diamantini et al., 2018) and according to the "objects", to which they refer (Sawadogo et al., 2019b).

In the first classification, Diamantini et al. (2018) distinguish three categories of metadata, which, according to them, are not independent of each other, but on the contrary - have intersecting points:

- **Business** metadata, which applies a business context to datasets and may include: descriptions related to the content and its use; owners and integrity restrictions; tags and properties to create a taxonomy on the collected data sets. Such data is usually entered by business users at the data ingestion stage (Sawadogo & Darmont, 2021).
- **Operational** metadata, which includes information generated automatically during data processing. This contains descriptions of the source and target data, such as location data, file size, number of records, data quality, identifier of the processes that created or transformed the data, status of the processes on the data, etc.
- **Technical** metadata, which includes information about the format and schema of the data. This is data that relates to the physical aspects of data sources and the application of data access policies based on defined attributes. Modern systems and tools supporting metadata management processes can search the file system or databases that are data sources and automatically discover possible dataset candidates and their proposed set of attributes (George, 2023). These crawlers not only look at sources and metadata, but also scan the data and apply machine learning models to identify sensitive information and suggest appropriate tagging.

The second classification divides metadata into 3 main categories – intra-object metadata, inter-object metadata and global metadata (Sawadogo et al., 2019b):

- **Intra-object metadata** present a set of characteristics associated with individual objects in the lake. This includes information

about properties, updates (versions) and data transformation, as well as summary metadata. To this group we also refer the semantic metadata, which are annotations that help understand the meaning of the data and are useful for discovering the relationships between objects (Hai et al., 2016).

- **Inter-object metadata** describe the relationships between data and are grouped into different categories depending on the scope of the data, its origin, its logical combination and content similarity. This means that they provide context and description that helps understand the data and its meaning in the business environment.
- **Global metadata** applies to the entire data lake rather than specific datasets and includes semantic resources that relate to knowledge bases, indexes that represent data structures such as text dataset keywords, image patterns or colours, and logs that are used to track user interactions in the data lake, such as logging in, viewing, or modifying records (Sawadogo & Darmont, 2021).

Although the data lake implements the “schema-on-read” concept, to ensure proper integration, understanding and quality of the data, it is necessary to use some kind of data model (Hai et al., 2016). Such data modelling usually consists of a conceptual model that should be flexible, facilitate frequent changes, and therefore should not impose a fixed schema (Khine & Wang, 2017; Mathis, 2017). The necessary metadata can be collected by extracting information that is predefined, for example, by reading the header information or the metadata can be additionally extracted from the source along with the original raw data. In addition, data can be continuously enriched with metadata during its life in the data lake, for example, by identifying relationships between different data sets (Mathis, 2017) or by tracking the origin of the data information. These particularities in the organization and the extraction of metadata in data lakes inevitably lead to certain challenges in **their use in BI**.

Metadata management in data lakes is still an under-researched area that requires further research and the application of a comprehensive approach, tailored to the specifics of the variety, structuredness and volume of collected data. This, in turn, is associated with a number of challenges in terms of their storage, integration and use for future analyses. Without properly collected metadata, the data lake is difficult to use because the structure and semantics of the data remain unknown. In this regard, based on the studies conducted in the existing literature, we can deduce some key challenges in metadata management in DL:

- A key element of metadata management is the **quality of collected metadata**, which determines the quality of the data description, which, in turn, is directly related to its visibility and ease of use. Therefore, it is essential that data lakes have defined policies and rules for quality control. The most common data quality rules are: completeness, data type, scope, format, selectivity, cardinality, and referential integrity. One of the most popular approaches to assessing data quality is known as “data profiling”. It analyses various aspects of the data to extract various statistics and characteristics. This includes evaluating parameters such as the number of missing values (data completeness), the number of unique values (cardinality), and the percentage of unique values (selectivity). It also checks for data types, scope and format, as well as data integrity checks. Metadata management enables data lineage tracking - showing the path of data from its origin to its current state. This transparency enables effective impact analysis. Understanding how changes to data elements can impact downstream processes helps maintain data quality by predicting and mitigating potential issues. Metadata management helps standardize data definitions, formats and classifications. Standardization contributes to data consistency by reducing errors and inconsistencies across the organization.
- Data in organizations comes from many different sources, such as customer log files, financial and accounting reports, emails, social media platforms, company-specific software, cloud platforms, etc. Therefore, **the integration** of such huge in volume and diverse in structure raw data is a challenging task (Karadi, 2014). In this regard, we believe that in order to effectively manage a large volume of heterogeneous metadata, it is important to build such an architectural framework in the organization that allows the support of access to the multitude of current and possible future data sources. If the majority of data is coming into the DL in real-time or near real-time through CRM, cloud applications and customer feedback, then a solution that allows the integration to Hadoop, Spark and NoSQL repositories would be suitable.
- Data ingestion is the most discussed phase of metadata extraction (Sawadogo et al., 2019b; Hai et al., 2016). However, the information extracted during **the data processing and access phases** also has significant importance for the overall business analysis. Currently, existing metadata management tools support finding and understanding, but not provisioning and accessing data (Eichler et al., 2021). Based on our research, we believe that there is still a lack of a well-established comprehensive metadata management

approach that focuses on all data types (with varying degrees of structuredness) throughout the DL data lifecycle.

At the time of data ingestion, data managers should encourage users to “tag” new data sources or tables with detailed information about them. It would be good for this annotation ethic to be seen as a company-wide commitment to all incoming data. In this regard, data managers could require that all new records in the data lake be annotated and, over time, stimulate this collaborative culture. This would improve the discoverability, accessibility and, at the same time, efficiency of use.

The presented challenges lead to a need to develop a structured and systematic approach to the collection, documentation, maintenance and use of metadata in data lakes to guide organizations in dealing with the complex process of metadata management.

3. Conceptual model for managing metadata in data lakes

Data in DL does not have a well-defined scheme (Derakhshannia et al., 2020). The organization, management and processing of metadata in DL largely depends on the needs and characteristics of the organization, the requirements, the environment and the particular systems used. Different models for metadata management in DL are found in scientific literature and it should be noted that some authors do not fully disclose their solutions.

In some scientific studies (Sawadogo & Darmont, 2021), metadata management models in DL are divided into two main groups:

- **Graph-based** models that represent metadata as a graph with nodes and edges. These models facilitate the visualization and understanding of metadata interactions. They typically describe the relationships between: data and their metadata, different datasets showing their interactions and dependencies, different versions of metadata showing the history of changes, users and their access rights to different data. Overall, the graph structure makes the model extremely flexible and convenient for representing complex relationships and interactions between metadata in the data lake.
- **Data vault** models, which are based on the data modelling concept developed by Dan Linstedt (n.d.). Data Vault was created based on the existing CMMI, Six Sigma, TQM, SDLC, and Function Point Analysis methodologies (Linstedt & Olschimke, 2015). It is designed to organize a flexible and scalable data structure that aims to handle the complexity of

integrating data from different sources and ensure easy access and management of information.

Other researchers consider models according to their functions in the architecture, depending on which layer of the DL architecture they are intended for (Hai et al., 2021). They emphasize that the main challenges in metadata management are related to:

- the ingestion layer, as it is responsible for importing data from disparate sources into the DL repository;
- the processing of the ingested raw data in the maintenance layer and transforming it into a form suitable for queries or analyses.

One of the most popular models proposed in 2016 which provides a generic approach to metadata management that is flexible, extensible and applicable in different contexts is the Generic and Extensible Metadata Management System (GEMMS) (Quix et al., 2016). GEMMS is an abstract metadata management framework that enables the use of graphs or other structures to allow easy association and analysis of metadata. One of the key features of GEMMS is its scalability. Organizations can add new metadata types, attributes and functionalities to the model according to their specific needs and business rules. The initial concept of the model lacks the capabilities of data versioning and a connection management mechanism.

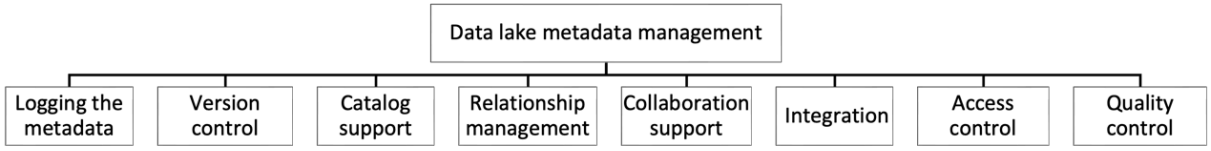
Another famous model is Ground (Hellerstein et al., 2017). It provides an abstract framework based on 3 main aspects related to metadata: Applications, Behaviour and Change. The Applications aspect concerns the specific applications or systems that use the data. It includes information about how the data is used, what its requirements are, and the context in which it is used. Behaviour includes information about the interaction and history of data, its versions over time, and its dependencies. Change refers to the changes that occur in the data and its context. The model records the interactions and operations related to the creation, updating, access between users and the data lake, but does not allow storing multiple representations of the same data, or the so-called data polymorphism.

Sawadogo et al. (2019b) carry out a study of the models in the considered area and as a result propose a metadata MEDAL model, based on the concept of objects (such as tables, columns, collections) and the relationships connecting them. It is implemented through a typology of metadata in the three categories that were discussed in section 3.1. – intra-object, inter-object and global metadata. As an evolution of MEDAL, the goldMedal method, created by a team of French scientists, can be considered. It is based on four main concepts: data entity, grouping, linking

and processing, which are defined at the conceptual and logical levels (Scholly et al., 2021).

A well-known graph model for representing metadata is HANDLE. It is designed to allow the addition of new functionalities and modules, making it flexible. The model maintains versioning of the metadata, allowing tracking of changes to it over time (Eichler et al., 2020). CoreKG (Knowledge Graph Management System) is another general and scalable graph-based approach for managing large amounts of knowledge and the relationships between them (Beheshti et al., 2023). CoreKG is designed to be scalable, allowing it to handle large and complex knowledge graphs with billions of vertices and edges. The lack of support for metadata versions can be cited as a disadvantage of the model. Another model that emphasizes scalability is EMEMODL (Cherradi & Haddadi, 2023).

The review of existing metadata management models in the data lake gives us a reason to summarize that the researched frameworks mainly cover the processes of metadata registration, metadata registry maintenance, data change tracking, and data-metadata relationships. These are the main ones important for the organization and management of metadata processes, but we believe that for metadata management in DL to be effective, more emphasis should also be placed on operations related to monitoring and tracking data quality as well as on comprehensively covering the processes of managing the access to metadata and defining the rights to edit, view, add metadata. In addition, emphasis should also be placed on the possibilities for cooperation between users and the creation of an environment for sharing, commenting and annotating metadata. Figure 2 presents a summary of all the important metadata management functions.

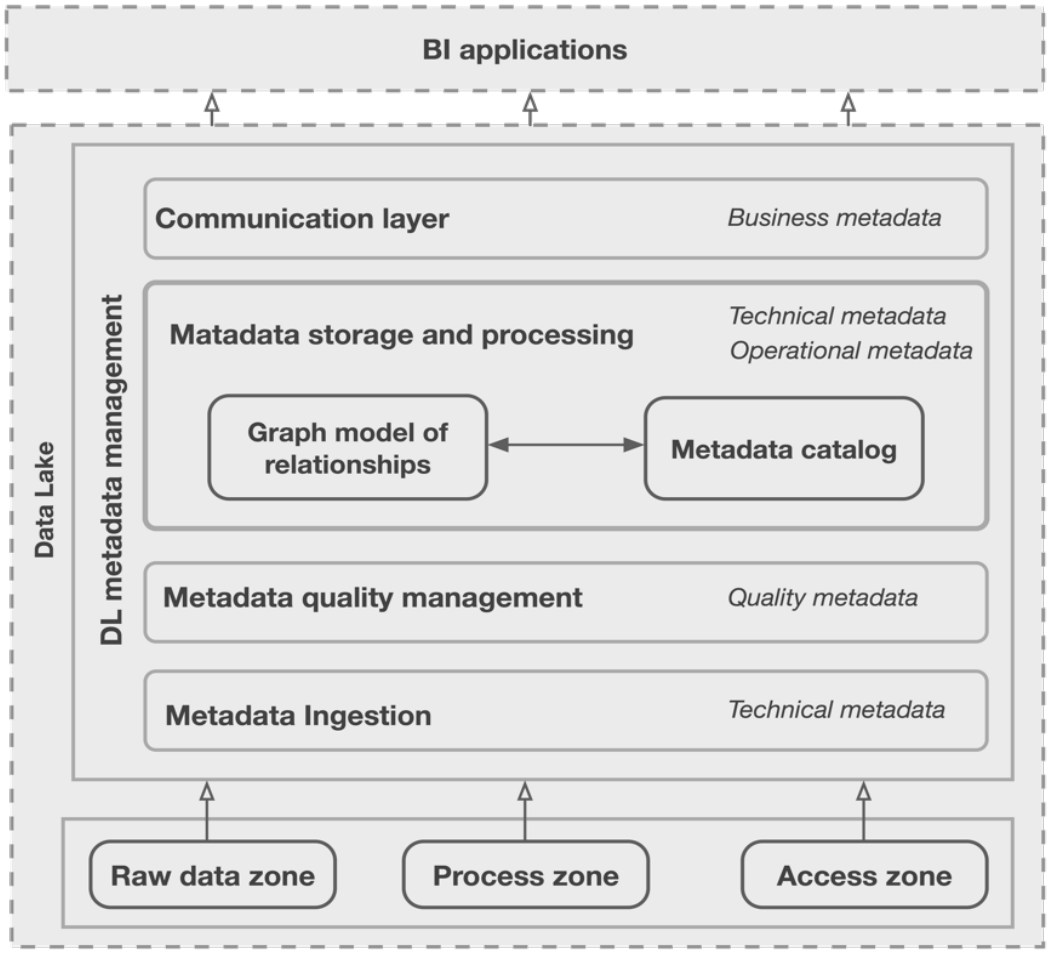


Source: own elaboration

Figure 2. Data lake metadata management functions

It is important to note that current research leads us to agree with Franck and Yan's view that there is no general metadata management system that works on heterogeneous data throughout its life cycle (Ravat & Zhao, 2019). It has been found that most of the discussed methods mostly cover the processes related to the Ingestion layer of the DL, and post-

ingestion modelling and metadata extraction is less affected. The authors of the current study believe that metadata management during data processing and transformation is essential to prevent data lakes from becoming incomprehensible data swamps. In order to improve the understanding of the data, as well as its use by the various analytical applications, the necessity of improving the processes of metadata organization and management should be considered. Therefore, a conceptual framework for the organization and management of metadata has been proposed (Figure 3).



Source: own elaboration

Figure 3. A conceptual framework for the organization and management of data lake metadata

The main logical layers of the conceptual framework are:

● **metadata ingestion** – responsible for identifying data related to business operations, automatically generated data from IoT devices, server log files, etc., as well as for generated data of Internet origin – social media

posts, emails, web content. Technical metadata is typically applied at this layer to help understand data formats, storage locations, and source details, which facilitates proper data ingestion and processing.

- **metadata storage and processing** – maintenance of a catalog with metadata and a model of the relationships between the data. Building the metadata catalog begins with the data entering the repository and continues as techniques are applied to process it. This layer makes extensive use of operational metadata to track the history of the origin and processing of the data, what the business rules associated with an object and other data objects are. It helps in monitoring data transformations and maintaining data quality throughout its lifecycle.
- **metadata quality management –** managing the accuracy of metadata records, provenance, accessibility, the completeness of the records. Poor metadata quality can lead to ambiguity, poor recall, and poor search and analytics application performance. At this layer, the use of so-called quality metadata is essential. Quality metadata is "information about the quality level of stored data in organization databases, and is measured along different dimensions such as accuracy, currency, and completeness" (Moges et al., 2016). High-quality metadata encompasses measurements and metrics related to data quality, which can involve assessing dataset status, data freshness, executed tests, and the outcomes of those tests.
- **communication layer** – refers to the processes of information exchange and instruction transmission between different elements of a metadata management system. In the communication layer, business metadata is used, often to help data users in understanding the business context of the data. Such data may include ownership details, policies for the use of the data, including any restrictions and metadata identifying whether the data adheres to specific regulations or standards, such as GDPR or HIPAA compliance.

The proposed conceptual framework represents a model for metadata management in organizations. It can be adapted to different types of data stores, both data lakes implementations and data warehouse and lakehouse architectures and can be used to build a metadata management strategy. When applying it, it is recommended to follow these steps:

1. Assessing the current state of metadata management in the repository. Outlining current issues and challenges in metadata management. Defining the goals that need to be achieved through more effective metadata management.

2. Analysing incoming data, their formats and volume. Defining the metadata to be managed (e.g., data types, sources, structure, data quality, etc.) and developing connectors or integration modules to connect to various data sources in the data lake.

3. Designing a comprehensive metadata model that captures relevant information about the data within the data lake. Defining metadata attributes such as data types, relationships, ownership, quality metrics, and usage history and developing specific rules for entering or generating metadata.

4. Selecting an appropriate metadata management software that has integration capabilities with existing systems and data lake architecture, facilitating automated metadata collection, scanning, parsing, and profiling of data.

5. Providing mechanisms for users to manually enrich metadata by adding additional information, such as business vocabulary terms, data classifications, and custom annotations.

6. Providing security measures to control the access to metadata based on user roles and permissions. Ensuring that sensitive metadata is protected and that access is granted only to authorized individuals.

7. Developing connectors or plugins for popular BI tools to ensure the successful integration of the metadata management framework into the BI workflow. This may include support for standard interfaces such as ODBC, JDBC, or REST APIs. Seamless integration allows BI users to work efficiently within their familiar tools, which accelerates learning.

8. Performing periodic analysis and optimizing the metadata management processes after registering changes in business requirements and the technological environment.

9. Implementing monitoring tools to track the performance and use of the proposed metadata management framework. Creating procedures for ongoing maintenance, including updates, debugging, and scalability adjustments.

From a software perspective, the proposed metadata management framework can be implemented using multiple platforms and data processing technologies. To ensure the adaptability and future-proofing of the framework according to the dynamics in technological and business needs, the following conditions should be taken into account for a smooth implementation:

- it must support different metadata storage standards;
- it must have capabilities for integration and interaction with the relational and non-relational databases used;

● it should offer support for modern interfaces and protocols, such as RESTful API, GraphQL, SOAP, and others, to facilitate integration with various systems;

● it must be modular and extensible so that it can adapt to new standards and technologies with minimal effort;

● to be able to integrate with business intelligence systems;

● to enable monitoring of activities;

● continuous feedback loops with users and stakeholders need to be established to capture evolving business requirements.

By considering these aspects, the aim of this research is to present a metadata management framework that not only meets current needs, but also provides the ability to adapt to future technological changes and evolving business landscapes. As an example of suitable software tools, we can point to one of the most popular platforms for DL according to current research – Apache Hadoop (Benjelloun et al., 2023; Gorelik, 2019). Within the Apache Hadoop ecosystem, there are tools that can be used to manage metadata in the Data Lake.

4. Discussion

Big data can provide competitive advantage to organizations only if it is properly collected and identified (Peicheva, 2021; Stoyanova & Vasilev, 2020). For the organization and management of big data, in many cases it is appropriate to use the DL concept. It is based on a number of applications in the field of AI (Armiyanova & Aleksandrova, 2022). Metadata is important to DL because it serves to unify and integrate data from different sources, facilitate the search and selection of the right data for BI analyses, maintain and manage data in warehouses and provide information about data quality, provenance and usage.

In the data lake, metadata extraction is essential to access datasets at a later stage (Hai et al., 2023). The lack of a unified view of metadata increases the difficulty of data management. Therefore, researchers believe that the proposed conceptual framework for metadata management in the data lake is important for the successful implementation of BI analytics and for the implementation of intelligent solutions in organizations. It provides the necessary foundation and structure to optimize the BI processes and improve data quality and analytical capability. It is based on good practices and provides

guidance for the organization and management of metadata in repositories with heterogeneous data sources.

The proposed conceptual framework and approach to its application are essential for business analysis and intelligent decision making, because with properly maintained metadata, data analysts can easily find and understand the data they need. In the context of applying artificial intelligence and machine learning technologies, proper metadata management improves data quality and model performance. In contrast to existing models discussed in Section 3, our conceptual framework not only comprehensively covers all examined aspects of the metadata management process but also provides a depth that surpasses the key aspects of metadata management addressed in the literature. From a practical point of view, its advantages are related to the possibility of its implementation and use in different types of modern data warehouses, both in those that are based only on the data lake concept, and combined ones that use the data warehouse and data lake concepts.

Organizations depending on their size and the sector, in which they operate, have different business requirements, regulatory standards, and technology infrastructures. The proposed conceptual framework is suitable for organizations that use DL repositories. It can be applied both to large organizations that have repositories using the integration of a data warehouse, DL concepts, and to smaller organizations that do not have a data warehouse in place, but use DL storage and integration with structured, operational data.

The proposed model is a good basis for building a metadata management architecture and supports the implementation of analysis and business intelligence strategies of companies. However, it has its limitations as it does not fully cover various aspects related to planning and organizing the data lifecycle, including issues related to data maturity. Also, subject to future development is the discovery of additional capabilities for automatic metadata mining. The directions for further research in the context of the article refer to some key areas that need research and development. Future work will be related to the implementation of the framework in specific organizations, which will be related to its detailing and defining precise guidelines for each of its layers.

Conclusions

Although the importance of metadata management in DL is recognized and affirmed in the scientific literature and practice, there is still a lack of clarity

about the necessary set of tasks related to metadata processing and storing which must be performed to prevent the risk of turning them into useless and obscure data. Therefore, in this paper, a clearer and comprehensive metadata management model is proposed. It provides valuable guidance to overcome some of the existing challenges in the research area.

Some of the key contributions of the presented framework are the provision of a comprehensive concept for the storage and integrated access to metadata by maintaining specific metadata catalogs, a graph-based model for representing the relationships between metadata, as well as defining specific quality management policies and rules of metadata. The authors of the current study believe that if organizations make efforts to implement tools that facilitate collaboration between all users using the data in data lakes by annotating, categorizing and sharing the metadata, the quality of the metadata will be significantly improved, and hence the efficiency of the entire process of their management. It seems that metadata is a kind of bridge between raw data and analytical processes, enriching information with context, semantics and structure.

Therefore, it can be concluded that well-organized and managed metadata in data lakes provides an intellectual structure that supports business analytics through improved data interpretation, faster access and discovery of valuable information, improved data quality assessment, avoidance of duplication and improving integration processes between different data sources.

References

Alserafi, A., Abelló, A., Romero, O., & Calders, T. (2016). Towards Information Profiling: Data Lake Content Metadata Management. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pp.178–185. doi:10.1109/ICDMW.2016.0033

Armbrust, M., et al. (2020). Delta lake. *Proceedings of the VLDB Endowment*, *13*(12), pp.3411–3424. doi:10.14778/3415478.3415560

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In *Proceedings of CIDR* (Vol. 8).

Armyanova, M., & Aleksandrova, Y. (2022). Artificial Intelligence System Problems and Opportunities to Solve Them with Design Patterns. *Izvestia Journal of the Union of Scientists - Varna.*

Economic Sciences Series, *11*(2), pp.172–183. Available online: https://ideas.repec.org/a/vra/journl/v11y2022i2p172-183.html

Bankov, B. (2018). An Approach for Clustering Social Media Text Messages, Retrieved from Continuous Data Streams. *Science. Business. Society: International Scientific Journal*, *3*(1), pp.6–9. Available online: https://stumejournals.com/journals/sbs/2018/1/6

Benjelloun, S., Aissi, M. E. M. E., Lakhrissi, Y., & Ali, S. E. H. B. (2023). Data Lake Architecture for smart fish Farming Data-Driven Strategy. *Applied System Innovation*, *6*(1), 8. doi:10.3390/asi6010008

Chen, Z. (2022). Observations and expectations on recent developments of data lakes. *Procedia Computer Science*, *214*, pp.405–411. doi: 10.1016/j.procs.2022.11.192

Cherradi, M., El Haddadi, A. (2023). EMEMODL: Extensible Metadata Model for Big Data lakes. *International Journal of Intelligent Engineering and Systems, 16*(3), pp.231–243. doi:10.22266/ijies2023.0630.18

Chihoub, H., Madera, C., Quix, C., & Hai, R. (2020). Architecture of Data Lakes. *Data Lakes, Volume 2*, pp.21-39. doi:10.1002/ 9781119720430.ch2

Couto, J., Borges, O. T., Ruiz, D. D., Marczak, S., & Prikladnicki, R. (2019). A Mapping Study about Data Lakes: An Improved Definition and Possible Architectures. *Proceedings*. doi:10.18293/seke2019-129

Cristescu, M. P., Mara, D. A., Culda, L. C., Nerișanu, R. A., Bârã, A., & Oprea, S. (2023). The Impact of Data Science Solutions on the Company Turnover. *Information*, *14*(10), 573. doi:10.3390/ info14100573

Derakhshannia, M., Gervet, C., Hajj-Hassan, H., Laurent, A., & Martin, A. (2020). Data Lake Governance: towards a systemic and natural ecosystem analogy. *Future Internet*, *12*(8), 126. doi:10.3390/fi12080126

Diamantini, C., Lo Giudice, P., Musarella, L., Potena, D., Storti, E., & Ursino, D. (2018). A new metadata model to uniformly handle heterogeneous data lake sources. In *Communications in computer and information science*, pp. 165–177. doi:10.1007/978-3-030-00063-9_17

Eichler, R., Giebler, C., Gröger, C., Schwarz, H., & Mitschang, B. (2020). HANDLE - a generic metadata model for data lakes. In *Lecture Notes in Computer Science*, pp. 73–88. doi:10.1007/978-3-030-59065-9_7

Eichler, R., Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2021). Enterprise-Wide metadata Management. *Business Information Systems*, pp.269–279. doi:10.52825/bis.v1i.47

Francia, M., Gallinucci, E., Golfarelli, M., Leoni, A. G., Rizzi, S., & Santolini, N. (2021). Making data platforms smarter with MOSES. *Future Generation Computer Systems*, *125*, pp.299–313. doi:10.1016/j.future.2021.06.031

Gorelik, A. (2019). *The Enterprise Big Data Lake*. Boston: O'Reilly.

Hai, R., Geisler, S., & Quix, C. (2016). Constance. *In Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16)*. doi:10.1145/2882903.2899389.

Hai, R., Quix, C., & Jarke, M. (2021). *Data lake concept and systems: a survey. arXiv (Cornell University)*. Available online: http://export.arxiv.org/pdf/2106.09592

Hai, R., Koutras, C., Quix, C., & Jarke, M. (2023). Data Lakes: A survey of functions and systems. *IEEE Transactions on Knowledge and Data Engineering*, *35*(12), pp.12571-12590. doi:10.1109/tkde.2023.3270101

Hellerstein, J. M., Sreekanti, V., Gonzalez, J. E., Dalton, J., Dey, A., Nag, S., Ramachandran, K., Arora, S., Bhattacharyya, A., Das, S., Donsky, M., Fierro, G., She, C., Steinbach, C., Subramanian, V. R., & Sun, E. (2017). Ground: a data context service. *Conference on Innovative Data Systems Research*. Available online: http://cidrdb.org/cidr2017/papers/p111-hellerstein-cidr17.pdf

Holom, R., Rafetseder, K., Kritzinger, S., & Sehrschön, H. (2020). Metadata management in a big data infrastructure. *Procedia Manufacturing*, *42*, pp.375–382. doi:10.1016/j.promfg.2020.02.060

Johnson, A. (2023, March 28). *How to use data lakes to improve data visualization*. Available online: https://datafloq.com/read/use-data-lakes-to-improve-data-visualization-2

Kadadi, A., Agrawal, R., Nyamful, C., & Atiq, R. (2014). Challenges of data integration and interoperability in big data. *2014 IEEE International Conference on Big Data (Big Data)*, pp.38–40. doi:10.1109/BigData.2014.700448

Khine, P., & Wang, Z. (2017). A New Ideology in Big Data Era. *In: 4th International Conference on Wireless Communication and Sensor Network (WCSN 2017), Wuhan, China*, *17*.

George, L. (2022, November 22). *Technical vs. Business Metadata: The Importance of Metadata Management.* Available online: https://www.okera.com/blogs/technical-vs-business-metadata-management/

Linstedt, D. (n.d.). *Data Vault Series 1 – Data Vault Overview*. Available online: https://tdan.com/data-vault-series-1-data-vault-overview/5054

Mathis, C. (2017). Data lakes. *Datenbank-spektrum*, *17*(3), pp.289–293. doi:10.1007/s13222-017-0272-7

Megdiche, I., Ravat, F., & Zhao, Y. (2021). Metadata Management on Data Processing in Data Lakes. In T. Bureš, R. Dondi, J. Gamper, G. Guerrini, T. Jurdziński, C. Pahl, P. W. H. Wong (Eds.), *SOFSEM 2021: Theory and Practice of Computer Science,* pp.553–562.

Metadata Management Tools Market Size, share & Trends analysis Report by metadata type (Business, Technical, Operational), by deployment (Cloud, on-premise), by application, by end-user, by region, and segment Forecasts, 2022-2030. Available online: https://www.grandviewresearch.com/industry-analysis/metadata-management-tools-market-report/

Miloslavskaya, N., & Tolstoy, A. (2016). Big data, fast data and data lake concepts. *Procedia Computer Science*, *88*, pp.300–305. doi:10.1016/j.procs.2016.07.439

Moges, H., Van Vlasselaer, V., Lemahieu, W., & Baesens, B. (2016). Determining the use of data quality metadata (DQM) for decision making purposes and its impact on decision outcomes — An exploratory study. *Decision Support Systems*, *83*, pp.32–46. doi:10.1016/j.dss.2015.12.006

Nambiar, A., & Mundra, D. (2022). An overview of data warehouse and data lake in modern enterprise data management. *Big Data and Cognitive Computing*, *6*(4), 132. doi:10.3390/bdcc6040132

Nogueira, I., Romdhane, M., & Darmont, J. (2018). Modeling Data Lake Metadata with a Data Vault. *22nd International Database Engineering & Applications Symposium (IDEAS 2018)*. doi:/10.1145/3216122.3216130

Peicheva, M. (2021). Artificial Intelligence in Human Resource Functions – Nature and Practical Application, Journal "*Choveshki resursi & Tehnologii = HR & Technologies*", Creative Space Association, 1, pp.3–12.

Prabhune, A., Stotzka, R., Sakharkar, V., Hesser, J., & Gertz, M. (2017). MetaStore: an adaptive metadata management framework for heterogeneous metadata models. *Distributed and Parallel Databases*, *36*(1), pp.153–194. doi:10.1007/s10619-017-7210-4

Quix, C., Hai, R., & Vatov, I. (2016). Metadata Extraction and Management in Data LakesWith GEMMS. *Complex Systems Informatics and Modeling Quarterly*, *9*, pp.67–83. doi:10.7250/csimq.2016-9.04

Ravat, F., Zhao, Y. (2019). Data Lakes: Trends and Perspectives. *In Proceedings of the International Conference on Database and Expert Systems Applications; Elsevier B.V.*, 11706 LNCS, pp.304-313.

Sawadogo, P. et al. (2019a). *Metadata management for textual documents in data lakes*. Available online: https://arxiv.org/abs/1905.04037

Sawadogo, P., Scholly, É., Favre, C., Ferey, É., Loudcher, S., & Darmont, J. (2019b). Metadata Systems for Data lakes: Models and features. In *Communications in computer and information science*, pp. 440–451. doi:10.1007/978-3-030-30278-8_43

Sawadogo, P., & Darmont, J. (2020). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, *56*(1), pp.97–120. doi:10.1007/s10844-020-00608-7

Scholly, E. (2021, March 24). *Coining GoldMEDAL: A new contribution to Data Lake Generic Metadata Modelling*. Available online: https://arxiv.org/abs/2103.13155

Skluzacek, T. J., Kumar, R., Chard, R., Harrison, G., Beckman, P. G., Chard, K., & Foster, I. (2018). Skluma: An Extensible Metadata Extraction Pipeline for Disorganized Data. *2018 IEEE 14th International Conference on e-Science.* doi:10.1109/ escience.2018.00040

Stoyanova, M., Vasilev, J., & Cristescu, M. P. (2021). Big data in property management. *AIP Conference Proceedings*. doi: 10.1063/5.0041902

ISSN 0861-6604
ISSN 2534-8396

D. A. TSENOV ACADEMY OF ECONOMICS

BUSINESS management

Published by the D. A. Tsenov Academy
of Economics – Svishtov

SVISHTOV * 2024 YEAR XXXIV * BOOK 2

The printing of the issue 1-2024 is funded with a grand from the Scientific Research Fund, Contract KP-06-NP5/42/30.11.2023 by the competition "Bulgarian Scientific Periodicals - 2024".

Submitted for publishinq on 07.06.2024, published on 10.06.2024,
format 70x100/16, total print 80

BUSINESS management

D. A. Tsenov Academy of Economics, Svishtov

Year XXXIV * Book 2, 2024

CONTENTS