JEL: C8

A CONCEPTUAL MODEL FOR THE ORGANIZATION AND STORAGE OF METADATA FOR DATA FROM INTERNET SOURCES

Snezhana Sulova[1]

[1] Assoc. Prof., PhD, Informatics Department, University of Economics – Varna, Bulgaria.
E-mail: ssulova@ue-varna.bg

Abstract

Data from internet sources, as well as their processing methods, is a prominent area of study. The evolution of web services and concepts that have developed with the growth of the Internet, as well as the widespread use of the world wide web and the wide variety of sensors and mobile devices have led to the Internet becoming a rich data source. In order to process and analyze this data, modern approaches and architectural solutions for their management are needed. The heterogeneous nature and specificity of Internet data requires the use of repositories, where the data can be organized and stored based on a pre-established management concept. In this regard, this article offers a model for organizing and storing metadata to help successfully manage data extracted from Internet sources. Such a model is a good basis for building data management architecture from heterogeneous sources that will support companies in analyzing and improving their business intelligence strategies.

Keywords: metadata; data; internet sources; data lake

Introduction

The development of Internet technologies and standards is one of the reasons for the continuous increase in data volume. According to the Internet Data Corporation (IDC), the amount of data that was created and replicated in 2020 was 64.2 ZB, which was an abnormally larger amount than previous years. Analysts forecast that the amount of data that will be created over the next 5 years will be twice as much as the data that has been created since the creation of digital data storage. The study also suggests that the data from the Internet of Things (IoT) is the fastest growing segment of data, followed by social media data (IDC, 2021). According to Bankov (2017), the interest in the processing and application of mined data has grown exponentially in recent years due to the increase in digital information that has been published on social media sites. The wide variety of sensors and mobile devices have turned the Internet into a rich data source. In addition the data has increased because of the many documents that have been digitized and the increase in people working, training and meeting on virtual platforms. Channels of Internet-based communication have helped companies digitize their physical methods of communications which has also allowed companies to succeed in the globalization and virtual organization of their businesses.

The increased amount of information on the Internet has become a useful additional source of new data for businesses and after appropriate processing can be turned into valuable additional information and data constituting a strategic resource that can be used by companies to make management decisions. The data from user interactions and business interactions from the world wide web come in various forms, structures and generation dynamics. Most often the data is hypertexts, other types of documents, multimedia content, structures of interaction, etc.

The ever-increasing digital information requires more complex storage, processing and protection (Armiyanova, 2020). For the storage and organization of different types of data, and if necessary, the extraction and transformation of it, a number of concepts for flexible data storage have been created such as Data Warehouse (DW), Data Lake (DL), and Lakehouse. No matter what data management scheme one chooses, a necessary condition for the successful use of the data repository is the creation of an integrated framework for its management. In order to help the building process of a modern data management architecture, this article would like to suggest a conceptual model for the organization and storage of metadata, which would help to successfully manage the repositories of heterogeneous data that has been extracted from internet sources.

1. Modern Repositories for Data Extracted from Internet Sources

Data, extracted from the Internet, has a heterogeneous character which is dynamic and often changes. The data’s volume is variable and cannot be determined unambiguously. In some cases, it can be relatively small, while in other cases it tends to increase rapidly.

Depending on its source, data that is extracted from the internet can be:

- Data from web-based and mobile applications such as applications for electronic trading, reservation systems and other web-based applications. This type of data is structured and stored in databases.
- Data from applications operating with IoT technology. This data is generated from real-time streams and has different characteristics, structured – most often with different numeric values and unstructured – videos, audio recordings, satellites and other types of images.
- Data from forums, blogs and other websites – comments, product reviews, user click streams; website structures.
- Data from social media – publications, comments, pictures, text messages, connections.
- Data from email messages – emails and their content.
- Data from web applications and sites usage – site visit files, server log files.

It is known from data management theory that in order for data to become information and then business intelligence, it needs to be organized, stored and transformed in the proper way. Storing data in various databases began in the 1960s. Software aimed at storing and retrieving user data is a database management system (Kuyumdzhiev and Nacheva, 2019). The necessity for more in-depth data analysis is the reason for the creation of Data Warehouse (DW). Traditionally, DWs have been used to integrate large amounts of corporate data into a single repository. Bill Inmon defines DWs as “a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management’s decision-making process” (Inmon, 2002, pg. 31). They have numerous benefits such as helping users save time as well as improve the decision-making process. However, it should be noted that their creation requires time and resources for modeling and preparing the data.

In a DW, data is selected and extracted from company databases then transformed and reformatted so that it can be stored (Curtis and Cobham, 2008, pg. 247). The appearance of so many, varied data sources makes it difficult to quickly transform, integrate and store the data. Traditional “schema-on-write” approaches are gradually changing allowing for “schema-on-read” principles, which provide for the storage of new types of unstructured data, to be applied.

A new concept of data storage, Data Lake, is being introduced. The main idea behind a Data Lake is to create a repository where data from different sources regardless of format can be stored (Laplante and Sharma, 2016). The DL concept also allows for data storage from internet sources such as social networks, devices using IoT concepts and other types of corporate data. The data is easily accessible and can be restructured, summarized and transformed by various applications.

The main advantages of the DL concept are:

- the storage of different types of data – structured, semi-structured, and unstructured;
- the ability to store data in an unprocessed form, which will allow for it to be converted whenever necessary;
- provides one location for the storage of data from different sources;
- allows for the use of various instruments to extract and process the data.

There are also a number of challenges associated with DL such as:

- the data may not be of good quality, due to the fact that it is received without supervision, management and pre-transformation;
- the performance of data operations is not guaranteed;
- the danger of the data turning into a "swamp" if it is not stored in an ordered way and organized by themes or categories and if it does not maintain the proper metadata.

Modern working conditions require the use of large volumes of corporate data as well as unstructured data extracted from web pages, web applications and other Internet sources. Therefore, DWs and DLs must be used in a complementary fashion. A group of scientists from Databricks and from the universities of Berkeley and Stanford, studying modern data architectures, are proposing to extend the functions of DLs in order to transform them into a high-performance system that can provide functions for managing data warehouses. This type of system design is called Lakehouse (Armbrust, Ghodsi, Xin, Zaharia, 2021). The main idea of a Lakehouse is to combine the key advantages of data lakes and data warehouses. The authors note that Lakehouses are particularly suited for working in cloud environments with completely separate computing nodes.

Whether the built-in architecture of the data is based on DL or the simultaneous use of the DW and DL concepts, the main problem continues to be the need for the integrated use of structured, semi-structured and unstructured data. Working with data from different sources requires the creation of flexible schemes, good engineering and a synchronization between systems. Developing a strategy to deal with data streams would be a huge advantage (Stoyanova, Vasilev, Cristescu, 2021). In the case of data warehouses lacking an explicit scheme or description exists the risk that those warehouses will turn into data swamps where data will be difficult to find and therefore become useless for the user.

2. Metadata and its importance for the organization and storage of data

The challenge of combining and storing multiple diverse data sources is mainly connected with the management of metadata. Metadata’s goal is to describe, explain and localize data in order to facilitate the data’s extraction and use. Pomerantzas (2015) defines metadata as a statement about a potentially informative objective. Inmon (2010) describes it as “glue”, which connects the different working components of data repositories.

The increasing availability of diverse data has led to the necessary appearance of different methods for storing metadata. Metadata also known as “data about data” is generated upon its entry into a data repository where it is converted and managed. It can take various forms. Good summaries and systematization of different types of metadata in Data Lakes have been accomplished by Benaissa et al. (2020). The authors have categorized the types of metadata into three groups which are represented in Figure 1.

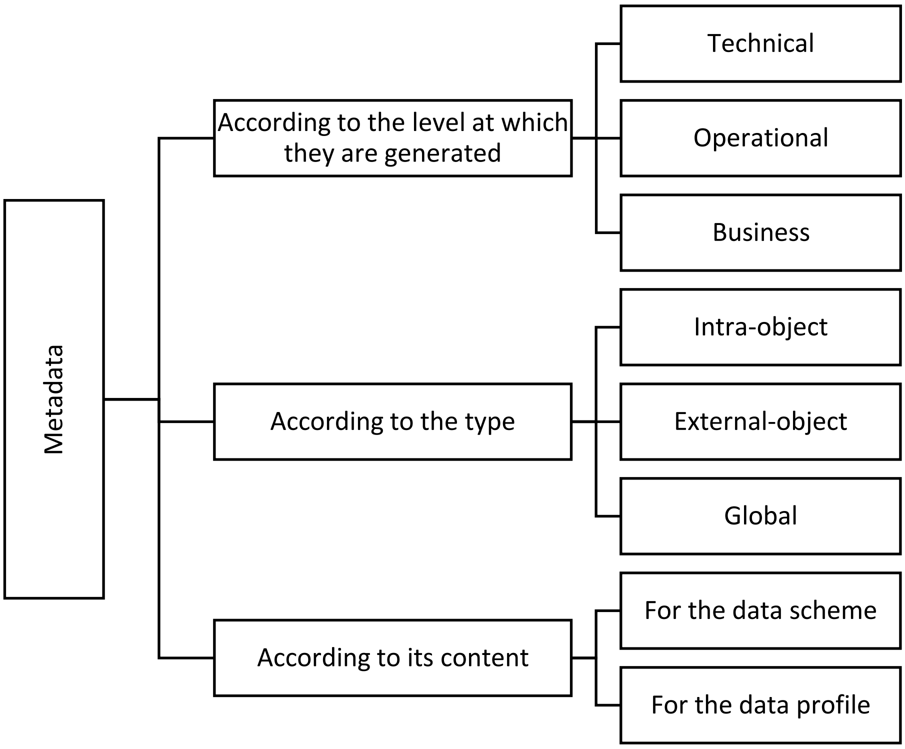

Fig. 1. Types of Metadata (Benaissa et al. (2020)

Metadata is often differentiated according to when it is generated, when it enters the data repository or during the time of its processing and use. Based on the level at which the data is generated, it can be:

- **Technical metadata** – includes the type and format of the data (text, image, JSON, etc.) and the structure of the data.
- **Operational metadata** – information, generated automatically at the time of data processing, e.g. for quality, origin, tasks executed, etc.
- **Business metadata** – for users, data use, descriptions, business rules, field restrictions, etc.

Metadata can be determined as belonging to a specific object, linking objects or for the whole repository and can therefore be defined as: **intra-object**, **external object** and **global** (Sawadogo, Kibata, Darmont, 2019).

According to its content, metadata can be seen as a **data scheme**, such as a number of attributes, attribute names, data types and **data profiles**, which describe data sets.

The availability of information concerning the origin of the data helps specialists identify where the data has come from, its format and what it is being utilized. The process of data integration and migration is facilitated when one has the full and accurate data information, knows its life cycle and how it has been transformed. Documented rules on the quality of data are important for assessing its ability to be processed. The fact is that software systems are becoming more and more complex and in order to work optimally and maintain the processes, it is essential to have suitable metadata documentation available.

Metadata is an information resource that must be managed (Haynes, 2018). The proper management of metadata can contribute to building repositories that allow for the integrated use of structured and unstructured data in business analytical processes. It should be noted that a major role is played by the specialists in this process through their knowledge, skills and competencies (Marinova, 2016). Allen and Cervo point out that present-day businesses have a lot of costs associated with managing duplicate information as well as developing and maintaining unnecessary software systems for processing low quality data. The authors specify that well-organized and well-maintained metadata is crucial for the efficiency and success of the various ways that data is managed and analyzed.

3. A Conceptual model for the organization and storage of metadata for heterogeneous data repositories

Even though the concept of DLs is relatively new, different studies have suggested models for the organization of metadata. Some authors emphasize the creation of a metadata management system that is focused on working with text documents in data lakes (Sawadogo, P., Kibata, T., Darmont, J. (2019). Others have focused on creating a metadata model designed for a medical data lake (Eder and Shekhovtsov, 2021). French researchers have developed a data management model with a comprehensive data lake system consisting of text documents and spreadsheets data (Sawadogo, Darmont, Nous, 2021).

The ever increasing resources as well as the diversity of sources on the world wide web have created challenges related to finding a data management method based on well-organized metadata systems. Due to the specificity of the data from internet sources, we consider metadata management to be a core function for modern data repositories and we would like to present a metadata management model for working with and organizing both structured and unstructured data throughout its life cycle.

Data management, based on metadata, must be created based on the following policies and processes:

- The publication and maintenance of a catalog of metadata, which has been collected at the moment the data has entered the repository and after the implementation of monitoring processes.
- Defining flexible rules for the configuration and management of the access to the data in a data repository. Not all incoming data is the same, some data must be managed in a precise, high quality manner while others may require less accuracy. This creates a need for different management rules.

The conceptual metadata management model for data repositories for internet sources is presented in Figure 2.

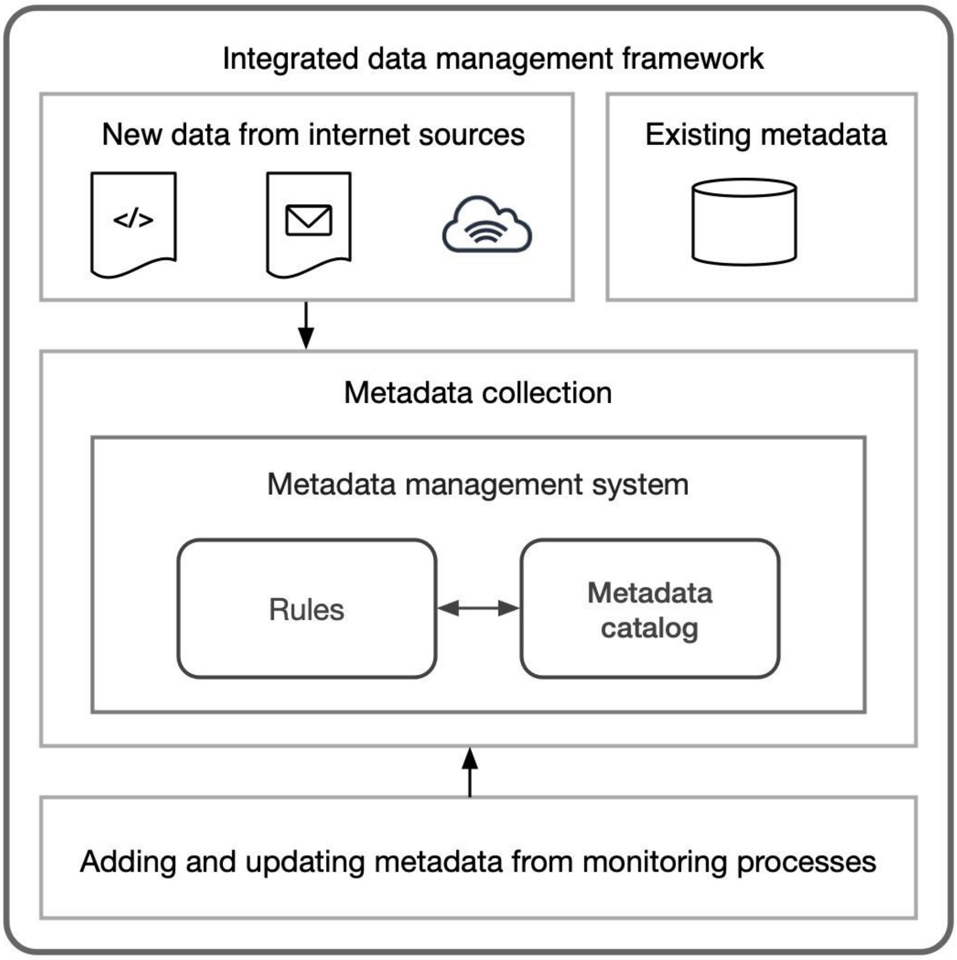
Integrated data management framework
New data from internet sources
</>
Existing metadata
Metadata collection
Metadata management system
Rules
Metadata catalog
Adding and updating metadata from monitoring processes

Fig. 2. A conceptual model for the organization and storage of metadata for data from internet sources

An essential element of the metadata management system is the **metadata catalog**. It is a kind of library that maintains a list of the metadata types and provides context that makes it possible to find and understand the corresponding sets of data. In addition, it stores all the changes that have occurred in the repository.

Based upon the research of data repositories for internet sources, we propose that the catalog has the following elements:

- **Profile characteristics of incoming data sets** – source; time of extraction; application which has created the resource; type, structure, format and resource size; text language; name that is given to the resource; with which elements it is connected; information about the resource rights;

- **Processing data** – access information, such as the names of users who have had access to the data sets and the access tools; life cycle information – transformation and data processing; data sensitivity which guarantees security; description of the various operating parameters, condition codes, etc.

The construction of a metadata catalog is a process, which starts with the entry of data into the repository and continues with the application of techniques for its processing. The accumulated data in the metadata catalog helps one to understand who has created the data, who is using the data, what are the business rules related to it, other objectives of the data, and the level of security. Since data repositories for internet sources represent a logical association of data structures with different organization schemes, it is important to store and update information about the location of data and its original source in the metadata catalog. The presence of a flag showing how the data has been entered into the repository and whether it has entered through streaming, another type of extraction or part of a web application database as well as its path after entry would help analysts discover the correct data and apply the appropriate management methods.

As we pointed out, working with data from Internet sources and using a DL presents a number of challenges, but we believe that the creation of the proposed concept for managing databases of metadata significantly simplifies the processes of finding and connecting information and could be the basis for analytical processes for organizations. The model can serve as a basic framework and specify a set architecture for the data by defining exact objectives for which the metadata will be stored and set rules for its storage and management. Defining the exact set of metadata and managing it for a specific data warehouse is a process of in-depth analysis. And as it is known, forming an accurate assessment in all projects is associated with the creation of a system of well-selected, defined, specific criteria and indicators (Delinov and Eskenazi, 2014).

Conclusions

Metadata extraction is one of the main features of modern data storage and management models. This article offers a concept for the creation of a model for the organization and storage of metadata that will help manage data extracted from internet sources. Such a model is a good foundation for building an architecture for managing data from internet sources and will assist in the implementation of analyses that will improve business intelligence strategies of companies. In addition, it would also allow for the application of flexible business analytical schemes that are built upon the idea of analyses adapted to the specific needs of a company and reacting to the quickly changing conditions of business.

In conclusion, it should be pointed out that the organization and storage of metadata is the basis for the development of modern architecture of data management. The proper design and development of models for the metadata management would assist in ensuring that the correct data is used at the right time using the right analytical process which would help the business intelligence strategies of companies.

References

1. Allen, M. and Cervo, D. (2015) *Multi-Domain Master Data Management. Advanced MDM and Data Governance in Practice*. Elsevier Inc.
2. Armbrust, M., Ghodsi, A., Xin, R., Zaharia, M. (2021) Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. *1th Annual Conference on Innovative Data Systems Research (CIDR '21),* Available at https://cs.stanford.edu/~matei/papers/2021/cidr_lakehouse.pdf, (Accessed 28 December 2021).
3. Armiyanova, M. (2020) Applying Patterns to E-Government. *Izvestia Journal of the Union of Scientists – Varna*. Economic Sciences Series, Varna: Union of Scientists - Varna, 9(1), pp. 156-167.
4. Bankov, B. (2017) Extracting Top Trends from Twitter Discussions in Bulgarian. *Izvestia Journal of the Union of Scientists – Varna*. Economic sciences series, 2, pp. 254-259.
5. Benaissa, R., Boussaid, O., Mokhtari, A., Benhammadi, F. A Comprehensive Study of Recent Metadata Models for Data Lake. *Data Analytics 2020, The Ninth International Conference on Data Analytics.* International Academy, Research, and Industry Association (IARIA), pp. 78-83.
6. Delinov, E. and Eskenazi, A. (2014) An Approach for a more Objective Evaluation of Practical Projects, Used in the Training Process, Serdica Journal of Computing, 8, Sofia, pp. 409-432

7. Eder, J. and Shekhovtsov, V. (2021) Data quality for federated medical data lakes *International Journal of Web Information Systems*. Emerald Publishing Limited 1744-0084, Vol. 17, No. 5, pp. 407-426.
8. Haynes, D. (2018). *Metadata for Information Management and Retrieval*. 2nd edn. Facet Publishing.
9. IDC (2021) *Data Creation and Replication Will Grow at a Faster Rate than Installed Storage Capacity, According to the IDC Global DataSphere and StorageSphere Forecasts*. Available at https://www.idc.com/getdoc.jsp?containerId=prUS47560321 (Accessed 26 December 2021).
10. Inmon, B. (2010) Data Warehousing 2.0. Modeling and Metadata Strategies for Next Generation Architectures. White Paper. Available at https://www.bitpipe.com/detail/RES/1327034955_829.html, (Accessed 27 December 2021).
11. Inmon, B. (2002) *Building the Data Warehouse. 3rd Ed.* Toronto: John Wiley & Sons, Inc.
12. Kuyumdzhiev, I. and Nacheva, R. (2019) Correlation Between Storage Device and Backup and Restore Efficiency in MS SQL Server. *Serdica Journal of Computing, Institute of Mathematics and Informatics*. BAS, 13, 3-4, pp. 139-154.
13. Laplante, A. and Sharma, B. (2016) *Architecting Data Lakes. Data Management Architectures for Advanced Business Use Cases*. O’Reilly Media, Inc.
14. Marinova, O. (2016) Business intelligence and data warehouse programs in higher education institutions: current status and recommendations for improvement. *Econimics and computer science:* Electronic journal, Issue 5, pp. 17-25.
15. Pomerantz, J. (2015) *Metadata*. Cambridge, The MIT Press.
16. Sawadogo P.N., Darmont J., Noûs C. (2021) Joint Management and Analysis of Textual Documents and Tabular Data Within the AUDAL Data Lake. In: *Bellatreche L., Dumas M., Karras P., Matulevičius R. (eds) Advances in Databases and Information Systems. ADBIS 2021.* Lecture Notes in Computer Science, vol 12843. Springer, Cham. https://doi.org/10.1007/978-3-030-82472-3_8.

17. Sawadogo, P., Kibata, T., Darmont, J. (2019) Metadata Management for Textual Documents in Data Lakes. *21st International Conference on Enterprise Information Systems (ICEIS 2019)*, May 2019, Heraklion, Greece. pp.72-83.
18. Stoyanova, M., Vasilev, J., Cristescu, M. (2021) Big Data in Property Management. Applications of Mathematics in Engineering and Economics: *Proceedings of the 46th Conference on Applications of Mathematics in Engineering and Economics (AMEE '20)*, 7 – 13 June 2020, Sofia, Bulgaria, Melville, NY: AIP Publ., Vol. 2333, № 1, 070001-1 - 070001-7.