Moroccan Data Lake Healthcare Analytics for Covid-19

Mohamed Cherradi, Anass El Haddadi and Hayat Routaib
Data science and competitive intelligence (DSCI), Applied Sciences Laboratory ENSAH/UAE, Al Hoceima, Morocco

Keywords: Data Lake, Metadata management, Big Data, Ontology, Covid-19.

Abstract: The coronavirus pandemic has radically changed the way we live. It has had a major impact on all areas. Researchers around the world are redoubling their efforts to explore solutions to this pandemic. In this context, many researchers attest to the driving role of technologies in the management and resolution of this global health crisis. To allow specialists in the field to better understand the key factors that influence the rapid spread of this epidemic, the establishment of a data lake information system is a major challenge for analyzing heterogeneous data to make effective decisions. This study aims to take advantage of the power of this massive amount of heterogeneous data to construct a data lake system as one of the expected information systems for many organizations. To meet this need, this paper proposes a new approach for data lakes based on dynamic ontologies to manage different data types regardless of their format. In addition to this, we have designed a friendly interface for the digitalization of hospital operations and a dashboard for visualization of the statistics of the covid19.

1 INTRODUCTION

Data Lake is one of the new concepts provided with the appearance of megadata. The original idea of data lakes came mainly from the business field rather than the academic field (Khine and Wang 2018). To confront the threats of big data and the shortcomings of data warehouses, James (Dixon 2010) gives birth to a new concept called data lake (DL). In this context, he said: “If you consider a Data Mart as a store of bottled water structured, cleaned, and packaged for easy consumption. The data lake is a large body of water in a more natural state”. This explanation proves the power of data lakes. But should not be examined as an approved definition. Data Lake is a nearly novel concept. Even though people consider it to be a marketing concept or strongly associate it with Apache Hadoop, which is not valid. There is no standard definition or architecture for data lakes; often, it depends on a use case (Ravat and Zhao 2019a).

Data lakes appeared as an adequate explanation to the problem of getting knowledge from a massive amount of diversified data sources (Diamantini et al, 2018). One of the good answers to this requirement is data lakes (Madera and Laurent 2016). A data lake is a central repository covering a vast amount of raw data in raw formats, separately from a data warehouse, a data lake based on a plane architecture. However, the management of the data stored in the lakes is assured by the presence of a well-maintained metadata layer. This allows us to understand, navigate, and extract knowledge in the end to make effective decisions.

The management of data lakes desires the definition of new techniques different from that adopted by conventional solutions, such as data warehouses (Diamantini et al, 2018). Most of these techniques are based on the idea of exploiting metadata which performs at its core and presents the essential tools granted to be a very aggressive plan in the big data era. Indeed, the data lake principles for each raw data need to be connected with a set of descriptive metadata. This represents the fundamental rule in data lake architecture because they make data possible actionable. For this purpose, the discovery of new models of metadata management depicts an open research issue in the data lake domain.

Automatic metadata extraction from different data sources is one of the significant challenges addressed in metadata management for data lakes. This consists of building an automatic metadata extraction process. Because the raw material of our approach is metadata, it's vital for data querying and navigation. Without any metadata management, the data lake appeared useless. In this context, we thought to automate this process via the development of a rest API.

Cherradi, M., El Haddadi, A. and Routaib, H.
Moroccan Data Lake Healthcare Analytics for Covid-19.
DOI: 10.5220/0010731700003101
In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML 2021)*, pages 232-238
ISBN: 978-989-758-559-3

There are several approaches and techniques in the literature to manage semi-structured and structured data ((Bhroithe et al. 2020), (Alloghani et al. 2019), (Aftab et al. 2020), (Ouaret et al. 2019)). However, it only focuses on two formats (structured and semi-structured) but does not examine unstructured data. In addition to that, most of the approaches that deal with unstructured data focus only on textual data (Yafooz and Fahad 2018).

The goal of this paper is formed as follows: Firstly, we take a look at the state of art and present a comprehensive vision of DL concepts. Secondly, we introduce a new method for structuring unstructured data. Especially in the health field, because in this field, we often find data in different formats. Thirdly, the construction of our ontology which represents our Moroccan data lake.

The remainder of this paper is structured like this: in Section 2, we parse the related literature. In Section 3, we present the formalization and data lake architectures adopted by our approach. Then we offer a procedure to partly structuring unstructured data sources. In Section 4, we describe our data lake ontology-based model to enrich the representation of unstructured data sources. In Section 5, we give an example case of covid-19. In Section 6, we present the evaluation technique and describe a critical discussion of our approach. Finally, in Section 7, we conclude our paper.

2 RELATED LITERATURE

2.1 State of the Art

Data lake relatively is a recent concept, introduced by James Dixon as an alternative to data marts; storing data into silos (Alserafi et al. 2016), to prevent them from being transformed into a data swamp must be accompanied by metadata (Sawadogo et al. 2019). The data lake model demands that any raw data be combined with a set of metadata. This represents a crucial competitive differentiator for any data lake architecture. Following (Farrugia et al. 2016), they proposed an approach to managing data lakes based on the extraction of metadata from an open-source data warehouse system named hive. To achieve their target, it applies Social Network Analysis techniques.

In the literature, various metadata classifications have been introduced. Thereby, various metadata models are used to design metadata classification. Among these models we find RDF. The power of this model is of course its semantic richness. However, its weakness is its complexity. Indeed, cannot maintain fast processing and analysis of the heterogeneous data.

A metadata model proposed by Oram is well-suited for data lakes (Oram 2016). There is also the model approved by Zaloni (Ben Sharma 2018), considered as one of the business managers in the data lake domain. Yet, Zaloni adopts a trinomial classification of metadata, namely operational, technical, and business metadata.

2.2 Data Lake Definition

A data lake is represented as an extensive system or repository that stores heterogeneous raw data; the diversification of concepts poses a significant issue. There is robust compliance in the literature on the definition of data lakes. Still, all existing definitions share the same vision about the definition of data lakes, respecting the idea that a data lake is a central repository of raw data stored in a natural format. For example, (Hai et al. 2016) defines data lakes as *"a megadata repository that stores data in its native format and provides on-demand ingestion functionality using metadata description"*. (Terrizzano et al. 2015) uses a definition of a data lake provided by (Madera and Laurent 2016) and asserts that "a data lake is a central repository containing enormous amounts of raw data described by metadata". Thus, we ascertain that there is a strong agreement concerning the definition of data lakes. In the situation of big data analytics, user needs are not established during the primary draft. A data lake is an answer that came with the appearance of big data, ingests raw data from different sources, and stocks source data in a natural format. Enables data to be processed conforming to diverse specifications. Indeed, empowers access to ready data for various needs, and supervises data to ensure data governance.

3 FORMALIZATION

In this section, we will describe our network model to manage the data lake, which will be used in our paper.

Our network model shows data lake as being a set of data sources, like this:

$$DL = \{DS_1, DS_2, \dots, DS_n \ / \ DS = \text{Data Source}\} \quad (1)$$

It is important to note that each data source DS_k is arranged with a set of metadata commented by M_k. We denote by M_{DL} the set of metadata repositories for heterogeneous data sources stored in the lake.

$$M_{DL} = \{M_1, M_2, \ldots, M_n \ / \ M = \text{metadata}\} \quad (2)$$

3.1 Typologie of Metadata

However, data comes from distinct sources, and various schemas are gathered together. It appears clearly that the metadata is essential to carry track of these data, playing a differentiator role in promoting the assistance of confusing data sources. This role is known as a vital interest in the promotion of data lakes; it is the only way to ensure effective management of data lakes so that they are transformed into data swamps in the absence of attribution of any scheme. Thereby, details on how to manage a metadata system are not provided. It remains an open research question. Each researcher can design their approach to build an efficient metadata management system that meets big data lake requirements.

Following what is said by (Oram 2016), metadata can be classified into three classes, i.e. Business metadata, which comprises business orders and gives data meaning in the context of the organization; Operational metadata, which contains the details of processing and accessing of data; Technical metadata, which includes information on the format and structure of the data, we often talk about data schema. Based on this argument, M_k can be expressed like the union of three sets, as you can see in figure 1.

Figure 1: Metadata classification (Diamantini et al. 2018).

Our network model is based on JSON, XML RDF, and many other semi-structured notations to represent metadata. Conforming to this notation, we designed our model, which is based on ontologies.

$$M = < N_i, A_i > \quad (3)$$

With $\mathbf{N_i}$, signify the node which represents a concept in ontology, identified by a URI, and $\mathbf{A_i}$: indicates an arc that represents the relation between the set of concepts belonging to the ontology.

3.2 Defining the Structure for Unstructured Data

Based on an extensible and generic ontology representation, our network model is fully matched for managing different types of data. The most significant difficulty concerns unstructured data, but the data lake often remains the leader for this type of requirement (Diamantini et al, 2018). A data lake is a powerful storage space with a large extent of raw data in native format. Thereby, for example, schema ingestion for unstructured data has no predefined schema, often we rely on metadata extraction to manage this type of data.

Our approach aims to contribute to this context. In particular, it proposes a REST API, able to automate the extraction of metadata from different data sources. Our methodology depends on using ontologies to represent all the data sources stored in the data lake utilizing an appropriate ontology. Indeed, ontologies represent one of the very flexible and on-demand solutions, often allowing the modeling of all use cases. In this article, we adopt networks to manage unstructured data, which is characterized by a frequently changing pattern. Moreover, we propose a technique for structuring unstructured data from keywords extracted from descriptive metadata. This is a fundamental task that represents a substantial asset because it facilitates the efficient management of non-structured data through our network model.

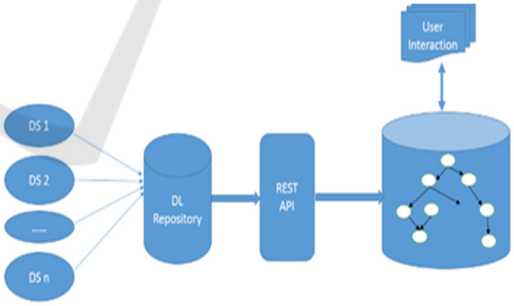

Figure 2: The architecture adopted by our model to structure unstructured data.

4 ONTOLOGY DATA LAKE ARCHITECTURE

There are two distinct models for data storage: the first one is a data warehouse, and the second one is a data lake. The question that arises is why data lake recently gained so much attention. In contrast, the storage market already has data warehouses for analyzing

business needs. Table 1 describes the contrasts between a data lake and a data warehouse. Nevertheless, if a Data Lake is a sandbox, it is still essential to give it an architecture. Otherwise, one day it ends up in the trash, and the data lake turns into a dark swamp made up of data that is forever inaccessible. In this context, data governance appears as one of the significant challenges of the proper functioning of a data lake (Panwar and Bhatnagar 2020). Furthermore, if we want to define what data governance is, we can say that it is the set of processes ensuring the management of corporate data sets. Allows us to apply policies, standards, practices, and strategies to manage data and to create value, trust. Governance ensures that trust data and that those responsible can be easily identified in the event of a problem. Poor governance is the source of many failures.

Table 1: Comparison between data lake and data warehouse (Simon Späti, 2018).

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
ETL, schema-on-write	PROCESSING	schema-on-read, ELT
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.

Data governance is a skilful blend of several skills: technologies, data science, digital marketing, and project management (Panwar and Bhatnagar 2020). Moreover, to sum up, data lake tends to meet the following challenges: not only data storage and processing. But also related skills such as visualization, data science, data governance, and real-time processing capacities.

As data lakes use a flat architecture, there are no processing tasks accomplished to accustom the structure to an enterprise schema. Each data source belonging to the data lake repository has a particular identifier and metadata properties to describe the data (as you see in table 2). However, data lakes don't need to follow the strict structure for controlling different types and aspects of data (Miloslavskaya and Tolstoy 2016).

Table 2: Data lake overview.

Data Sources	Identifier	Metadata Description
DS_1	ID_1	MD_1
DS_2	ID_2	MD_2
...	….	….
DS_n	ID_n	MD_n

For data lake architecture, the diversification of concepts proves a significant issue. Indeed, the architecture we have proposed in this paper (See Figure 2) forms planning of the management and storage data. It will optimize the information capacity of monitoring. It comprises various information pools, supporting an assortment of examination and mix instruments tools.

5 CASE STUDIES: HEALTHCARE ANALYTICS FOR COVID-19 IN MOROCCO

This section proposes an example case study to show the different stages of our approach based on ontology networks. Whose objective is to build a dedicated data lake for each hospital to follow the evolution of covid-19.

5.1 Digitalization of Hospital Operations

To follow the evolutions of covid-19 and to provide an interactive dashboard capable of representing and knowing the current state of our pay. In this context, it appears the digitalization of the hospital is a necessity of building a central data lake capable of storing and processing any type of data, and in particular in the case of the hospital, often we find heterogeneous data of different kinds are encountered.

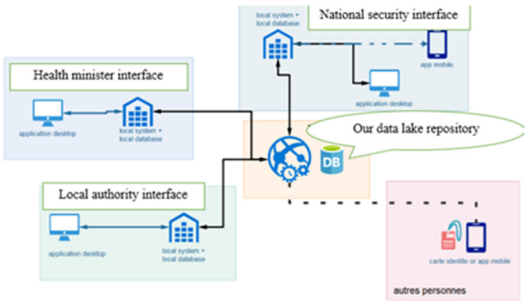

Figure 3: the citizen data collection process and their covid-19 tracking.

To make this evolution of citizens and to provide interactive visualizations. We have developed an interactive interface (see Figure 4), which allows the median to do CRUD operations. It allows to add, modify, delete in case of error and consult the citizens who carry the covid-19 virus.

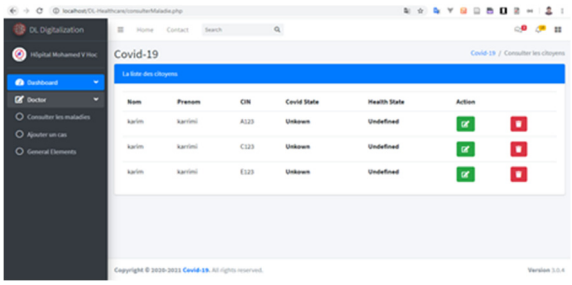

Figure 4: digitization operations interface

To improve decision-making, data visualization appears to be an effective solution, plays a significant role in governance and activity management; they allow managers of each service to identify correlations between objectives, performance and identify risk at a glance. In this context, we develop an interactive dashboard (see figure 5) about the citizens who carry the viruses. Via this dashboard, we can perform statistics in real-time on the situation of covid-19 in our country.

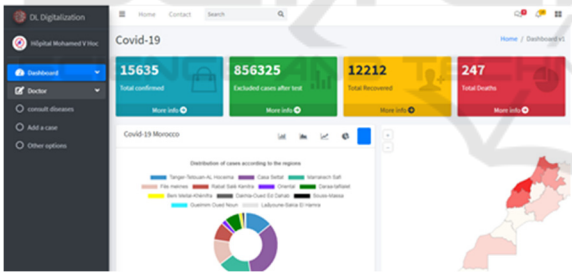

Figure 5: Dashboard visualization of covid-19.

5.2 An Example Case of Structuring Unstructured Data based Ontology

This section shows an example case study intended to explain the different stages of our network model. To realize our use cases, we designed our data lake with four unstructured data sources (i.e., two pdfs and two videos). All these sources store data about the health situation, such as the illness situation, the number of illnesses, the type of drugs being taken, etc.

Our approach follows the following steps to structure unstructured data in the lake, as we specified in Section 3.2. It is composed of two phases: Metadata Extraction and ontology representation. By applying the methodology referenced in Section 3.2, we attain the identical representations in our network-based model, shown in Figs. 6.

To visualize our ontology, we use the OntoGraph tool (resp. VOWL), which automatically generates our network model. And supports interactive navigation in the relationships of our ontology. OntoGraph and VOWL aim for an intuitive and complete representation that is understandable for users less familiar with ontologies.

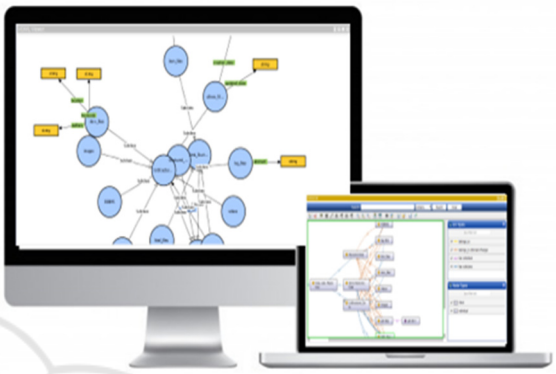

Figure 6: The ontology network corresponding to the data source.

6 DISCUSSION

This section is intended for a critical examination of our proposal based on several facets. It contains two sections. Firstly, we perform a comparison between our approach and the associated approaches. Secondly, we assess the performance of our approach.

6.1 Comparison between Our Approach and the Related Ones

If we start from the idea that we are going to use metadata to manage data lakes. In this section, we can classify three main categories of approaches most closely related to our own. These approaches aim more precisely to:
(i) Automate the extraction of metadata based on efficient solutions
(ii) Structure unstructured data
(iii) Extract knowledge from heterogeneous data
Regarding the first category, the corresponding approaches share the same objective with ours, i.e. automatic metadata extraction. However, most of the techniques that exist to automate the metadata extraction process (such as, for example, Metadata Extractor tool, Cermine, GROBID, Etc.) focus on a

well-specified data format. However, our technique is based on a rest API developed with Flask capable of extracting metadata from different data sources, whatever its structure, and this represents a competitive advantage compared to other techniques in this context.

The analogous methods manifest some similarities for the second category, but they contain some differences from ours. In particular, both start from the idea of exploiting metadata. However, most approaches only focus on textual data, while our approach maintains data sources with different formats.

For the third category, they extract knowledge from heterogeneous data sources. For example, the approach of (Lo Giudice et al. 2019) aims to query heterogeneous data collected in the lake. This approach has several distinctions from ours. Indeed, it is essentially based on the identity of the keywords extracted from the metadata for querying a data lake. However, our approach is based on semantic web techniques.

6.2 The Evaluation Criteria of Our Approach

The criteria for evaluating the achievement of our method to structure unorganized data consists of ascertaining in which sense can measure the connection between the concepts conforming to the keywords frequently applied to describe unstructured data. Given a network model similar to ours, an analytical technique of measuring performance is based on exploiting specific criteria generally approved in network analysis to measure the height of structuring. These measures are availability, relevance, consistency, update, and semantic precision.

In order to structure the unstructured data, our approach is based roughly on metadata. So in this context, to assess the performance of our approach, we must first measure the efficiency of metadata, as illustrated in Figure 7. All of them belong to the interval [0,1]; the more the value is close to 1, the structuration of data is perfect, and therefore, the model is magnificent in terms of structuring unstructured data.

Figure 7: measures to assess our approach

7 CONCLUSION

In this article, we have proposed a novel network model to effectively manage the data sources stored in the lake and avoid the famous dilemma of data lakes, which is known as data swamp. In addition to that, we defined a novel technique to extract complex knowledge from the data sources that exist in the lake based on dynamic ontologies. Finally, we measure the achievement of our approach to see the ability to structure unstructured data.

This paper should not be interpreted as finalized, but it should be seen as the starting point for a new set of efficient big data management techniques. Specially designed to manage heterogeneous data in the lake.

REFERENCES

Aftab Z, Iqbal W, Almustafa KM, Bukhari F, Abdullah M (2020) Automatic NoSQL to Relational Database Transformation with Dynamic Schema Mapping. In: Sci. Program. Accessed 5 FEB 2021 https://www.hindawi.com/journals/sp/2020/8813350/.

Alloghani M, Al-Jumeily D, Hussain A, Aljaaf A, Mustafina J, Khalaf M, Tan SY (2019) The XML and Semantic Web: A Systematic Review On Technologies. pp 92–102

Alqaryouti O, Khwileh H, Farouk TA, Nabhan A, Shaalan K (2018) Graph-Based Keyword Extraction. In: Studies in Computational Intelligence. pp 159–172

Alserafi A, Abelló A, Romero O, Calders T (2016) Towards Information Profiling: Data Lake Content Metadata Management. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW).

Ben Sharma (2018) Architecting Data Lakes, 2nd Edition - Ben Sharma - Google Livres. https://books.google.co.ma/books/about/Architecting_Data_Lakes_2nd_Edition.html?id=phFJzQEACAAJ&redir_esc=y. Accessed 10 Feb 2021

Bhróithe A, Heiden F, Schemmert A, Phan D, Hung L, Freiheit J, Fuchs-Kittowski F (2020) A Generic

Approach to Schema Evolution in Live Relational Databases. pp 105–118

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: ADBIS 2018 Short Papers and Workshops, AI*QA, BIGPMED, CSACDB, M2U, BigDataMAPS, ISTREND, DC, Budapest, Hungary, September, 2-5, 2018, Proceedings. pp 165–177

Dixon J (2010) Pentaho, Hadoop, and Data Lakes | James Dixon's Blog. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/. Accessed 10 Feb 2021

Farrugia A, Claxton R, Thompson S (2016) Towards social network analytics for understanding and managing enterprise data lakes. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp 1213–1220

Hai R, Geisler S, Quix C (2016) Constance: An Intelligent Data Lake System. pp 2097–2100

Khine PP, Wang ZS (2018) Data lake: a new ideology in big data era. ITM Web Conf 17:03025. https://doi.org/10.1051/itmconf/20181703025

Lo Giudice P, Musarella L, Sofo G, Ursino D (2019) An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake. Inf Sci 478:606–626. https://doi.org/10.1016/j.ins.2018.11.052

Madera C, Laurent A (2016) The next information architecture evolution: the data lake wave. pp 174–180

Miloslavskaya N, Tolstoy A (2016) Application of Big Data, Fast Data and Data Lake Concepts to Information Security Issues

Oram A (2016) Managing the Data Lake - Managing the Data Lake [Book]. https://www.oreilly.com/library/view/managing-the-data/9781492049876/titlepage01.html. Accessed 5 Feb 2021

Ouaret Z, Boukraâ D, Boussaid O, Chalal R (2019) AuMixDw: Towards an automated hybrid approach for building XML data warehouses. Data Knowl Eng 120. https://doi.org/10.1016/j.datak.2019.01.004

Panwar A, Bhatnagar V (2020) Data Lake Architecture: A New Repository for Data Engineer. Int J Organ Collect Intell 10:63–75. https://doi.org/10.4018/IJOCI.2020010104

Ravat F, Zhao Y (2019a) Data Lakes: Trends and Perspectives. In: Hartmann S, Küng J, Chakravarthy S, Anderst-Kotsis G, Tjoa AM, Khalil I (eds) Database and Expert Systems Applications. Springer International Publishing, Cham, pp 304–313

Ravat F, Zhao Y (2019b) Data Lakes: Trends and Perspectives

Sawadogo PN, Scholly É, Favre C, Ferey É, Loudcher S, Darmont J (2019) Metadata Systems for Data Lakes: Models and Features. In: Welzer T, Eder J, Podgorelec V, Wrembel R, Ivanović M, Gamper J, Morzy M, Tzouramanis T, Darmont J, Kamišalić Latifić A (eds) New Trends in Databases and Information Systems. Springer International Publishing, Cham, pp 440–451

Terrizzano IG, Schwarz P, Roth M, Colino JE (2015) Data Wrangling: The Challenging Journey from the Wild to the Lake. In: CIDR

Yafooz W, Fahad S (2018) Managing Unstructured Textual Data. 2:26–33