Improving a New Data Lake Architecture Design Based on Data Ponds and Multi-Agent Paradigms

Jabrane Kachaoui and Abdessamad Belangour

Abstract For several years, Big Data has been considered the next evolution for processing various data in Data Lake. However, one of the major difficulties for Big Data development is its need for structured knowledge. This is why, a large part of research focuses on ontologies formalization that has, among other things, given rise to ad hoc languages as Web Ontology Language (OWL). In this paper, ontologies were embedded into Data Ponds to operate and cooperate with agents. Indeed, a new architecture has been built based on traditional multi-agent system in a Big Data context. This architecture enables end users to access to information from Data Lake in real time with the slightest effort. This issue has been the subject of research for nearly ten years, but it remains one of the major obstacles to Big Data development.

Keywords Big Data · Data Lake · Multi-agent model · Machine learning · Ontologies · Data Ponds

1 Introduction

Companies put the capitalization of their data at the heart of their digital transformation to upgrade it later. Data Lake (DL) is positioned as one of solutions that enhance data capital of organizations. It is designed to catalog, capture, store, exploit, manipulate and manage a large amount of available data. First perceived only as low-cost storage environments [1], the potential for leveraging data it stores has transformed it into a strategic issue for organizations [2].

The Fig. 1 illustrates the potential market that this topic represents for organizations in world level and its importance in industrial world. The global DL market size is expected to grow from USD 7.9 billion in 2019 to USD 20.1 billion by 2024, at a Compound Annual Growth Rate (CAGR) of 20.6% during the forecast period.

J. Kachaoui (✉) · A. Belangour
Hassan II University, Casablanca, Morocco

M. Ben Ahmed et al. (eds.), *Innovations in Smart Cities Applications Volume 5*,
Lecture Notes in Networks and Systems 393,
https://doi.org/10.1007/978-3-030-94191-8_65

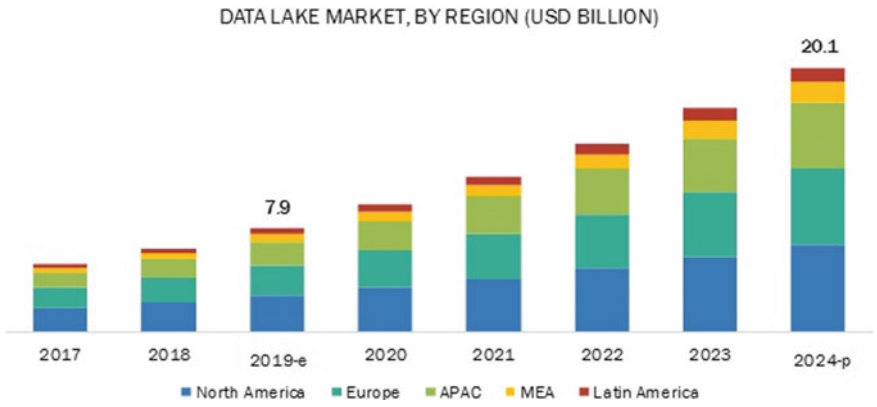

Fig. 1 Market development for Data Lake solutions (source: MarketsandMarkets Analysis)

The challenge of DLs is very real and the expectations of significant return on investment.

The adoption of DL concept has been accelerated with the convergence of the need for federated platforms in companies, to facilitate the exploitation of data, and technological possibilities around certain technologies associated dealing with large amounts of data [3].

The primary objective of DL is to allow exploration, without preconceptions, of all data that compose it, with a view to discovering new information referrals to be exploited in the context of an organization's data valuation. DL is a system driven by data that completes decision-making system in place. It becomes an essential component of information system, under the influence of Big Data [4, 5].

To better understand DL concept, its positioning in the information system, involves understanding its challenges at organization level and in industrial environment that sets it up. In the following sections, we present the synthesis of our research work on the challenges of DL and we propose a new architecture design to tackle these issues for a better data management.

The goal of this paper is to present our approach that improves Data Ponds (DP) data management by using agent paradigm in a new DL architecture. This paper is an extended work of our previous contribution in DL challenges [6], the complementarity of DL and Data Warehouse (DW) approaches based on a multi-criteria decision making method [7, 8]. New DL system using ontology model in Covid-19 use case [9], MQL2SQL (Mongo Query Language To Structured Query Language) for transforming data from MongoDB into RDBMS [10], a global proposal DL architecture design [11] and a clustering approach based on K-means based on metadata for DP construction [12]. A recent work focused on presenting ontology construction into DP [13].

In this paper, we develop the edge level, which collaborate, with our previously developed DL architecture in order to deploy the proposal architecture presented in this study that highlights the importance of DP in DL architecture construction.

The reminder of this paper is organized as follows. In section two, it presents the knowledge engineering work that has contributed to authors' reflection for this research. However, it is not intended to cover this domain in full because the literature offers this type of panorama. This contribution differs from traditional approaches, of which it ultimately contains few elements. However, this study will emphasizes on needs and results in terms of data analysis, because our goal is to build and maintain a new data management system. Section three describes validation and experimentation of the proposed architecture. Finally, section four concludes the paper and outlines future work.

2 Related Work

Our first work [6] on the state of knowledge on DL allowed us to list the main academic contributions around this subject, they are still limited but are increasing. As proof, the work of Ansari [14] around semantic profiling in DL. The work of Fang [15] is a first position in the literature on DL, for him, DL (1) supports data storage, in its native form, at low cost. This cost is low on two aspects, the storage is done on so-called "convenience" servers (X86 technology) and mainly because there is no formatting, cleaning or data preparation (step usually very expensive), (2) stores a wide variety of data types, from blobs to traditional DBMS, data multi-structured with multimedia data, (3) does not format data until it is used. This means that the efforts expensive modeling and data integration are reduced with this approach. This approach is known as schema on read, (4) performs analysis based on a single domain. As the value is initially fuzzy, users must therefore develop specific analysis to use data, (5) governance strategies are configured to identify, reuse and eliminate data. (6) Indicates data source, including the indications on their origin, who has changed them, what version of their change, etc.

For Fang [15], there is no particular basic architecture of DL and he strongly associates the creation of a DL with the implementation of an Apache Hadoop environment. In addition, he sees the decline of decision-making system in favor of DL, which he even sees in the long term in the cloud technology. In his work, the DL is seen as a data management methodology where all data of an organization is gathered, physically, on a platform based on Apache Hadoop.

He warns of the risk of data non-governance, like any conventional data management project. We deduce four limits to Fang's vision of DL:

- It focuses exclusively on Apache Hadoop technology.
- It does not take into account the fact that certain criteria could "block" the movement of data, such as data gravity.
- The governance axis is decoupled from DL,
- DL is seen as the "killer" of DW.

In [15], there is no proposition around the architecture of DL. As part of our works, we look to the industrial world to examine the architectures studied or implemented, and let's try to find their points of convergence and divergence to

establish our point of view. The work around architecture in industrial world is led by software vendors, related to Hadoop technology [16], such as Cloudera [17] or MapR [18] and their architecture vision is strongly influenced by their platform (Fig. 2).

IBM [19] is the only one to offer, through its work, a developed and mature architectural vision, certainly oriented by the software products that promotes, but which addresses a number of issues that we ask. The first works of IBM on DL does not use the word "Data Lake" but "Data Reservoir" [20], name that under pressure and marketing craze for the word "Data Lake" will be abandoned. However, this denomination of "Data Reservoir" is an indicator of the first industrial works which are beginning to question, quickly, Apache Hadoop technology as the only answer to DL implementation.

3 Towards a New Data Lake Architecture Design

The concept of DL was born from the Big Data and Apache Hadoop technology influence. Its design is adequate with current software on industrial market, and is therefore focused essentially on technologies. We have seen in previous section that the relation between DL and Apache Hadoop was very limiting and no longer corresponded to organizations expectations, as well as DL concept. We have seen its evolution in term of architecture with the appearance of hybrid architectures [21].

The idea behind the proposal architecture comes from the perception that in many domains, applications are not an isolated systems, but share needs, functionalities and properties. The general idea of this research is to take advantage of these commonalities to define a basic architecture from which new applications can be built easier, faster and with a higher level of quality.

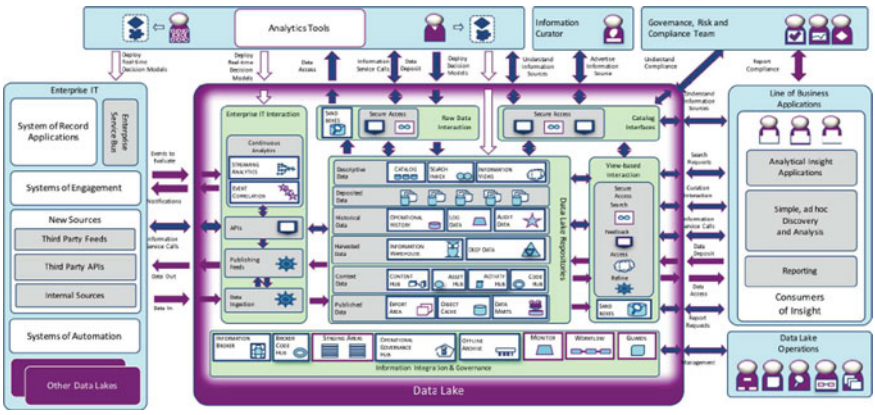

Fig. 2 Data Lake Reference Architecture proposed by IBM [19]

To date, there has been no work on the formalization of DLs in the scientific literature. The industrial software businesses posed the problem, essentially proposed software solutions but did not never addressed the formalization of DL [12].

As part of our work, we want to experiment this architecture based on agent paradigm, ontologies and assess the relevance of this approach for formalizing DL. Our expectations around this architecure are therefore (1) proposing a list of components to set up for operating DL without being transformed into Data Swamp. (2) Initiating a formalization process for DLs.

3.1 System Overview

The proposed system in this paper is introduced in Fig. 3. Global goal that agents and ontologies should achieve is to support process while using existing DW systems for business analysis and employing information from DP. To support this goal, this system includes several agent roles that are as following Server Agent (SA), Information Agent (IA). These agents have both reactive and proactive characteristics. Reactive are mainly due to responding to the environment according to the model defined in the Knowledge Base (KB) they use. Proactive are due to their ability to learn from the environment and change the initially defined KB to, for example, improve performance. Ontologies are used as a main interconnection object for domain knowledge representation, agent-to-agent communication and most important for agent-to-business user communication. An important element of an environment is DP, where agents acquire information for the purpose of decision making. Retrieved Information (RI) is saved in a KB and available for further employment for DW analysis. All information gathered from DP is considered by IA, where inference over several task ontologies used by individual agents (SA and IA) is performed. Moreover the sub goal of this system is delivering the right information at the right time to the right users. Business users are able to employ agents to perform tasks on their behalf. For example, managers in enterprises have to request reports from their systems – OLAP (Online Analytical Processing) or from transactional databases, and managers have to review reports every appointed period of time (day, week, month, etc.). This task of information acquisition is predecessor for decision-making and is more or less straightforward - business user sends a request for analysis and reviews the content according to some Key Performance Indicators (KPI). KPI is simply a measure of performance and is commonly used in enterprises to evaluate how successful they are. Tasks like this are automated and user participation is reduced as much as possible. An initial analysis model (e.g. OLAP or DM) has to be captured in the ontology by business users, while execution and optimization is left for agents. Business users first define initial parameters for analyses to be performed, while agents perform these analyses and recommend improvements. When some action is required from business user, he is notified and has the ability to act or change rules of agent's execution.

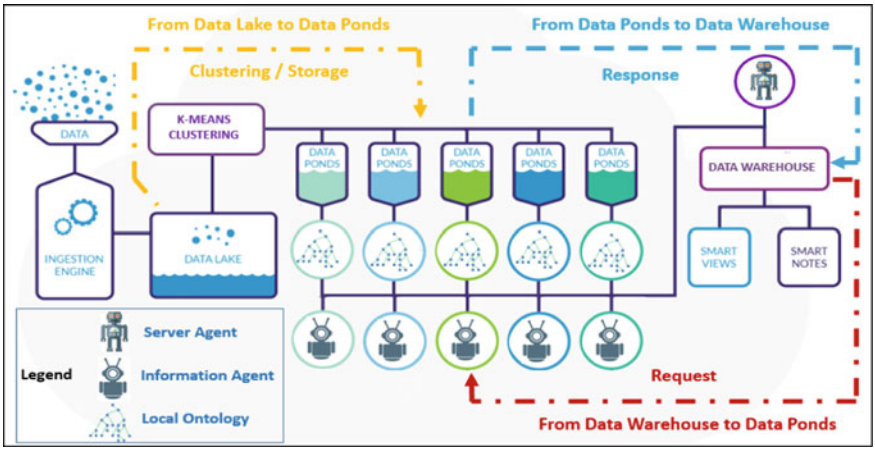

Fig. 3 System overview

To enable these functionalities we introduce ontologies as a mediation mechanism for knowledge exchange between actors (agents and business users) that cooperate in this system.

In this system, agents interact via communication languages (FIPAACL, KQML, etc.) in order to reveal information by sending messages to each other. In a message linked to a given communication language, information is represented in the form of a content expression, easy to transfer, encoded with the appropriate format (character string, sequence of bytes, etc.) and conforms to a content language, which may or may not be understandable by humans (SL, LEAP, etc.). However, each agent can have its own disposition to represent information, which is not always easy to handle, hence the notion of ontology that aims to facilitate the manipulation of messages exchanged between agents. These ontologies define a vocabulary and a set of relationships that connect different elements of this vocabulary, which represents an adequate solution to the proposed negotiation protocol, which goes beyond the limits of traditional agent communication.

4 Conclusion

The interest of our work for a wider scientific community lies first in having grouped together a number of standard methodologies for data processing within one system. Second it lies in the fact that this system describes a data workflow from the DL to end user from where its requests passed by an intelligent system of ontologies and agents that collaborate in order to meet companies needs in real time.

References

1. Gartner: Gartner says beware of the data lake fallacy (2014), http://www.gartner.com/newsroom/id/2809117
2. MarketsandMarkets: Data Lake Market worth $20.1 billion by 2024 (2024), http://www.marketsandmarkets.com/PressReleases/data-lakes.asp. Accessed June 2021
3. A. Panwar, V. Bhatnagar, Data Lake architecture: a new repository for data engineer. Int. J. Organ. Collective Intell. (IJOCI) (2020). https://doi.org/10.4018/IJOCI.2020010104
4. C. Madera, A. Laurent, The next information architecture evolution: the data lake wave. Int. Conf. Manage. Digit. EcoSyst. (2016). https://doi.org/10.1145/3012071.3012077
5. Cloudera. Turn Your Data Lake into an Enterprise Data Hub (2014), https://vision.cloudera.com/turn-your-data-lake-into-an-enterprise-data-hub/
6. J. Kachaoui, A. Belangour, Challenges and benefits of deploying big data storage solution, in *Proceedings of the New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society*, Article No. 22, (2019), pp 1–5
7. J. Kachaoui, A. Belangour, A multi-criteria group decision making method for big data storage selection, in *Proceedings of the International Conference on Networked Systems*, (2019), pp 381–386
8. J. Kachaoui, A. Belangour, An adaptive control approach for performance of big data storage systems, in *Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development*, (2019), pp 89–97
9. J. Kachaoui, et al., Towards an ontology proposal model in data lake for real-time COVID-19 prevention cases. Int. J. Online Biomed. Eng. (iJOE) (2020)
10. J. Kachaoui, A. Belangour, MQL2SQL: a proposal data transformation algorithm from MongoDB to RDBMS. Int. J. Adv. Trends Comput. Sci. Eng. **9**(2), 2457–2463 (2020)
11. J. Kachaoui, A. Belangour, From single architectural design to a reference conceptual meta-model: an intelligent data lake for new data insights. Int. J. Emerg. Trends Eng. Res. **8** (4), 1460–1465 (2020)
12. J. Kachaoui, A. Belangour, Enhanced data lake clustering design based on k-means algorithm. Int. J. Adv. Comput. Sci. Appl. **11**, 547–554, (2020) https://doi.org/10.14569/IJACSA.2020.0110472
13. J. Kachaoui, A. Belangour, Local ontologies merging in data ponds. Int. Conf. Intell. Syst. Comput. Vis. (ISCV) (2020). https://doi.org/10.1109/ISCV49265.2020.9204097
14. J.W. Ansari, Semantic Profiling in Data Lake. Ph.D. thesis, RWTH Aachen University, 2018
15. H. Fang, Managing data lakes in big data era : what's a data lake and why has it became popular in data management ecosystem, in *International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, (IEEE, 2015), pp. 820–824
16. K. Shvachko, et al., The Hadoop Distributed File System, IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST) (2010), https://doi.org/10.1109/MSST.2010.5496972
17. R. Menon, *Cloudera Administration HandBook*. (Packt Publishing Ltd., 2014)
18. MapR tutorial (2020). https://mapr.com/developer-portal/mapr-tutorials/
19. M. Chessell, F. Scheepers, N. Nguyen, R. van Kessel, R. van der Starre, Governing and managing big data for analytics and decision makers, (2014), http://www.redbooks.ibm.com/abstracts/redp5120.html?Open
20. P. Tyagi, H. Demirkan, (2016), http ://analytics-magazine.org
21. P.P. Khine, Z.S. Wang, Data lake: a new ideology in big data Era, in *WCSN 2017, Wuhan, China, ITM Web of Conferences*, vol. 17, (2017), pp. 1–6, https://doi.org/10.1051/itmconf/2018170302