ROYAUME DU MAROC
UNIVERSITE ABDELMALEK ESSAADI
FACULTE DES SCIENCES ET TECHNIQUES
TANGER

المملكة المغربية
جامعة عبد المالك السعدي
كلية العلوم والتقنيات
طنجة

CED : « Sciences et Techniques de l'Ingénieur »

AVIS DE SOUTENANCE

«MOHAMED CHERRADI»

Présentera ses travaux de recherche en vue de l'obtention du Doctorat en Sciences et Techniques

Intitulé de la thèse :

« Metadata Management for Data Lake Governance based on Machine Learning »

Date :	**Samedi 17 juin 2023**
Heure :	**10 heures**
Lieu :	**Amphi B, ENSA – Al Hoceima**

Devant le jury :

Membres de jury

Pr. Abdelhadi FENNAN	FST - Tanger	Président et Examinateur
Pr. Mohamed BELLOUKI	FP - Nador	Rapporteur
Pr. Mohamed BENATTOU	FS - Kénitra	Rapporteur
Pr. Jaber EL BOUHDIDI	ENSA- Tétouan	Rapporteur
Pr. Anouar Abdelhakim BOUDHIR	FST - Tanger	Examinateur
Pr. Mohamed BOUHORMA	FST - Tanger	Examinateur
Pr. Anass EL HADDADI	ENSA - Al Hoceima	Directeur de thèse

ABSTRACT

In the big data era, data is distinguished by its volume, variety, velocity, value, and veracity (5V). Beyond storage, the main difficulty with big data is getting high-quality value out of large, quick, and varied datasets. Lastly, the Data Lake (DL) has emerged as a fresh idea to tackle big data analytics issues. Although big data has been studied in several academic publications, but there are still many research drawbacks associated with it, such as the data variation. In fact, the variety of data sources frequently resides in "information silos," which are groups of decoupled data management systems with diverse data models, query languages, and schemas. With conventional "schema-on-write" systems like data warehouses, several issues are raised to efficiently integrate, access, and query the immense volume of different data in these information silos.

As a replacement of data warehouses for the storing and processing of massive data, the idea of a data lake has gained ground interest during the past ten years. To offer adaptable and scalable decision-support systems, data lakes use a schema-on-read methodology. A crucial challenge with data lakes is the requirement for a strong metadata system. Further, metadata are necessary to allow studies in the absence of a fixed data structure and keep the lake from becoming a useless data swamp.

Despite the fact that the literature appears to agree on the significance of metadata systems, there are still lots of unanswered questions and doubts around the implementation process. To organize metadata in data lakes, a number of methods have been put forth, however, most of them do not support industrialized analytics as data warehouses. Additionally, a sizable portion of the literature only allows data scientists access to the data lake, excluding business users. Further, the vast majority of current methods for managing metadata in data lakes only apply to structured and semi-structured data. Therefore, there is still need for research into how to design a metadata system that supports both industrialized analysis, which covers unstructured data.

Moreover, the difficulties of merging several diverse data sources in data lakes have their origins in the field of data integration research. However, the main tasks involved in data integration are understanding the links, such as schema mappings across data sources in lakes, and responding to user queries across diverse sources. Therefore, metadata management is essential, especially for accessing and querying the data, preventing a data lake from turning into a data swamp. Thereby, acquiring, modeling, storing, and enriching the metadata that characterizes the data sources are the key issues for metadata management in data lakes.

Additionally, to the best of our knowledge, there is no formal definition or architecture for the data lake because it is a relatively new concept. Regarding the scope of the context, the literature proposals are insufficient. Our initial contribution is consolidated into a comprehensive definition and a general design of the data lake that covers several zones, like ingestion, preparation, analysis, and the data governance zone. Thus, governance is essential to ensuring that data lakes are neither invisible nor unavailable to their diverse consumers. Then, a metadata management system is the key component of good governance. The approaches to managing metadata described in the literature are insufficient and not always applicable to DLs.

In this context, we suggest in this thesis a number of contributions to the literature on the conception and application of data lakes. Our contributions during this thesis are broken down into two parts:

1. Disambiguating the term "data lake" is the primary goal of our thesis. That is to say, at the outset of this thesis, data lakes were still rather new and poorly understood. In order to make this more clear, we suggest a new definition of data lakes and examine various methods for managing information and organizing the architectural structure of data lakes.

2. Based on a thorough state-of-the-art analysis, we pinpoint the advantages and disadvantages of current methods for managing metadata in data lakes, highlighting the fact that the majority of them are too specialized to be applied again and again. The few generic approaches are similarly constrained in terms of available capabilities and data types. Furthermore, to overcome these flaws, we propose in this thesis an extensive and scalable data lake architecture as well as a prototype system for metadata management, termed EMEMODL, which solves the issue of data accumulation without documentation and exploitation by offering data ingestion, integration, and querying over many data formats, including structured, semi-structured, and unstructured data. Our data lake solution enables the integration of many data storage systems with various data models to store heterogeneous raw data, whatever its formats. Additionally, our solutions provide an integrated querying interface by designing a novel query rewriting engine that incorporates descriptive mapping-based techniques for data integration with the big data processing system based on the DLDS approach and DLDB-Services, among other contributions. The mapping formalisms are applied to the computation complexity and decidability of some reasoning tasks using the well-known Grover quantum method. Further, our study proposal is separated into two primary stages in order to capture the schema of diverse data sources and provide clustering-based algorithms to detect link strength, which improve the metadata semantics and promote the quality of the data lake.

a. As a first stage, we suggested the following paradigm to control the complete data lifecycle in a DL:
   i. Metadata representing different types of ingested heterogeneous data and different ingestion modes, whether in batch mode or real-time.
   ii. Leveraging high-level procedures to specify metadata reflecting various data transformation processes.
   iii. To describe the analysis carried out in the DL, to be able to reuse, and to easily parameterize feature analysis, it is necessary to have metadata that are focused on analysis, and in particular machine learning.

b. In the second stage, we defined the metadata management system, which enables users to dynamically ingest metadata and intuitively explore the DL's content in order to reuse or modify them. Thus, the benefit of such a proposal is that it meets the demand for data science industrialization. Then, we have simultaneously carried out a performance analysis of metadata ingestion and a study examining the user experience to assess the viability and usability of our concept.

Such a framework for implementing a data lake, represents one of the most valuable contributions to the scientific literature. The results were encouraging and suggest that, under certain conditions, our proposals

might be highly helpful not only for this particular case study but also for any use case that wants to avoid misunderstandings and include data governance into its tasks.