Received: 16 December 2023 | Revised: 21 March 2024 | Accepted: 20 May 2024 | Published online: 27 May 2024

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2025, Vol. 3(1) 58–66
DOI: 10.47852/bonviewJDSIS42022312

BON VIEW PUBLISHING

Enhancing Data Lake Management Systems with LDA Approach

Mohamed Cherradi[1],* and Anass El Haddadi[1]

[1]Data Science and Competitive Intelligence Team, Abdelmalek Essaadi University, Morocco

Abstract: In today's fiercely competitive business landscape, data have emerged as a precious asset crucial for any company's growth. It embodies a genuine catalyst for economic and strategic advantages, distinguishing industry leaders from the rest. Prominent organizations recognize the importance of not just amassing data from diverse sources but also harnessing the transformative power of data analytics for informed determination processes. Within this setting, the data lake solution stands as a robust framework handling vast data sources and enabling data investigations to support decision-making tasks. This paper delves into the realm of intelligent data lake management systems designed to overcome the limitations of traditional business intelligence, which struggles to meet the demands of data-driven decision-making. Data lakes excel in the analysis of data from myriad sources, particularly when data cleaning becomes a time-consuming endeavor. Still, managing diverse datasets devoid of a predefined data structure presents a significant challenge, potentially leading to a data lake devolving into a data swamp. Within this article, we adopt the Latent Dirichlet Allocation model to oversee the data lake environment's handling, processing, analysis, and display of huge datasets. To evaluate the efficacy of our suggested approach, we conducted comprehensive assessments using the topic coherence metric. Our experiments yielded results indicating the superior accuracy of our approach when applied to the tested datasets.

Keywords: data lake management, Latent Dirichlet Allocation, data analytics, big data, topic modeling

1. Introduction

The idea of a "data lake" growing in favor as a practical way to handle vast and diverse data sources. It is evident that the adoption of big data technology yields substantial benefits for companies, driving remarkable improvements in their business operations. In this context, business intelligence emerges as a sophisticated strategy, transforming raw data into actionable insights [1]. Business intelligence not only uncovers fresh opportunities but also illuminates potential pitfalls, revealing novel business insights and refining the decision-making processes, thereby enhancing organizational productivity [2, 3]. Consequently, business intelligence has become a top priority across various industries. However, traditional business intelligence primarily focuses on structured data, overlooking the invaluable treasure trove of information concealed within unstructured data. This oversight results in an incomplete understanding of the environment, hampering effective decision-making. Moreover, conventional business intelligence systems excel in structured data analysis but fall short in managing unstructured data. The advent of big data introduces formidable challenges, considering that data are heterogeneous and come in a variety of structured and unstructured formats. Consequently, traditional business intelligence encounters severe limitations in fully harnessing the advantages of big data. Hence, a reevaluation of how we leverage the potential of business intelligence, from data ingestion to storage and analysis, becomes imperative. Therefore, the emergence of data lakes as a response to the challenges posed by big data presents a pivotal solution for managing both structured and unstructured data efficiently. Data lakes provide a unified platform for storing diverse types of data without the constraints of a predefined schema, enabling organizations to extract insights from the entirety of their data assets.

Further, the advent of the era of big data lakes has ushered in a new paradigm for data management, driven by the growing demand among companies to analyze a diverse array of data types. However, traditional solutions are struggling to keep pace with this evolving landscape. In response to the challenges posed by big data, innovative technologies like data lakes have emerged, providing organizations with the means to efficiently store and manage vast and heterogeneous datasets. Yet, the absence of a predefined data schema in data lakes necessitates robust metadata support to avoid the conversion of these lakes to a data quagmire, characterized by data that is either useless or lacks actionable insights, often referred to as undocumented data [4]. In this paper, we showcase an innovative application of the Latent Dirichlet Allocation (LDA) algorithm, demonstrating its unique ability to enhance the functionality of a flexible data store such as a corporation's data lake. Our approach introduces a novel perspective on data management, offering a streamlined solution to the challenge of handling heterogeneous data sources within the data lake environment. By harnessing LDA topic modeling, we provide organizations with a powerful tool to efficiently navigate through diverse data sets, mitigating the risk of data swamp and uncovering valuable insights. This novel integration of LDA

*__Corresponding author:__ Mohamed Cherradi, Data Science and Competitive Intelligence Team, Abdelmalek Essaadi University, Morocco. Email: m.cherradi@uae.ac.ma within the business data lake framework not only enhances data organization and accessibility but also promotes informed decision-making and strategic planning. Our research significantly contributes to the advancement of data management practices, offering a fresh approach to maximizing the utility of data lakes in data-centric landscape. Indeed, the business data lake's adoption approach simplifies the complex processing associated with traditional Enterprise Data warehouses (EDW), abbreviating their complexity. In direct comparison to a standard EDW, this approach not only enhances responsiveness to evolving business needs but also extends the operational lifespan of EDW systems.

Within the context of big data, the idea of data lakes has emerged as a complex and dynamic entity. Despite its innovative nature, data lakes encounter a multitude of challenges [5]. Nevertheless, research in the realm of data lakes is of utmost importance due to their unique capability to store heterogeneous data without predefined structures. In order to identify the main themes within the corpus of the data lake, this study presents the use of the LDA topic model as an efficient data exploration method. By identifying these advanced machine learning models, organizations gain the ability to tackle intricate issues, fathom concerns, interpret environmental feedback, gain competitive insights, and ultimately gain a strategic edge in delivering precisely tailored products or services to the right recipients at the opportune moments. This research contributes significantly to the exploration of data lake potential and its transformative impact.

In the dynamic landscape of data lakes, the ever-growing diversity and complexity of raw data formats present formidable challenges. This paper presents a pivotal contribution, centered on the exploration of abstracts within content documents, revealing latent themes that weave through an extensive corpus. We meticulously analyze the linguistic fabric of resources residing in the data lake, leveraging advanced topic modeling techniques, with a specific emphasis on harnessing the LDA model. By extracting concealed insights from this heterogeneity, we empower data consumers with the ability to navigate the vast data reservoirs within data lakes for analytical endeavors. This research directly addresses the imperative to streamline information discovery, retrieval, integration, and data optimization, which traditionally consumes over 70% of the resources in data analytics projects [6]. Our proposal offers a promising pathway to alleviate this pressing concern. Thus, topic modeling demonstrates efficacy in identifying a multitude of domains in the context of scientific publications. In light of this, this study examines strategies for efficiently retrieving, analyzing, organizing, and uncovering insights from text inputs. To put it briefly, topic modeling is an innovative and extremely successful technique for document classification automatically [7], comprehending vast amounts of textual data inside a large collection of unstructured documented data, and summarizing vast amounts of textual data.

The rest of this paper is structured as follows: Section 2 reviews previous research on the data lake concept and topic modeling approaches. Section 3 illustrates the suggested methodology. Next, Section 4 will handle the experiment's findings and analysis. Section 5 concludes our investigation and presents research perspectives.

2. Literature Review

In this section, we provide the vital backdrop for our study, equipping readers with the essential knowledge required to fully grasp the subsequent content of this paper. We commence by introducing the concept of data lakes and the intricate challenges they pose. Following this, we delve into a comprehensive review of the research landscape related to topic modeling, spanning its various applications and the evolving technologies that have shaped this field. This contextual groundwork forms a pivotal framework for a thorough comprehension of the ensuing sections in this paper.

2.1. Data lake

Ever since the industry first used the term "data lake," it has undergone a continuous evolution. Its origins can be traced back to its introduction by Dixon (Pentaho's CTO). Initially conceptualized to be a proficient remedy for managing unprocessed data, while accommodating diverse user requirements [8], the concept of a data lake stands in stark contrast to traditional data warehousing or data mart approaches, which necessitate rigorous data preprocessing and exploration. Unlike conventional data warehousing methods, data lakes offer the flexibility to circumvent the costly routine operations of data preparation by saving unprocessed data in its original format and keeping it "as-is."

Additional conversations on the compositions, functions, and applications of data lakes became more prominent. Notably, IBM took significant strides in addressing the matter of governance for data lakes, concurrently introducing the "data mess" concept aimed at rendering raw data consumable [9]. The process of data swamp entails a series of operations typically performed upon the creation of raw data within the data lake. These operations encompass the following: (a) selection of datasets relevant to business profitability; (b) exploration of the legal prerequisites associated with the chosen datasets; (c) direct loading of input data into data lake storage systems without undergoing expensive data transformations processes; (d) maintenance of the various data sources, incorporating detailed metadata that is both semantic and descriptive; (e) data preparation of analytical purpose; and (f) facilitation of diverse consumers' exploration, ultimately offering data visualization capabilities. Consequently, data lakes grapple with additional issues such as data integrity and protection. For instance, Marty [10] as well as Sitarska-Buba and Zygala [11] proposed a particular data entry methodology, emphasizing data security and event management. In this context, sensitive user data, particularly from human resource management systems, may not all be physically stored within the data lake; some remains within the original data repositories. This architectural approach accommodates the processing, analysis, storage, and querying of sparsely formatted personal data. Additionally, the data security significance has been underscored in proposals [12, 13] that underscore the vital role of safeguarding data in this evolving landscape.

Since the year 2016, data lakes have experienced an unprecedented surge in popularity, captivating the attention of both the business and academic communities. This burgeoning interest is reflected in the emergence of high-level proposals concerning data lake architecture [14–18], along with insightful comparisons with traditional data warehouses [19]. Academic publications have further enriched the discourse by exploring the fundamental concept, intricate components, and pertinent challenges associated with data lakes. Meanwhile, several leading companies have ventured into the realm of data lakes, offering commercial solutions to meet the evolving demands of the data landscape. Notable industry players such as Google, IBM, Microsoft Azure, AWS, Cloudera, Snowflake, Oracle, SAP, and others have introduced their own data lake offerings [20].

Furthermore, Teradata has contributed significantly to the data lake ecosystem by providing essential features via its open-source data lake development platform, including data intake, metadata management, and data governance [14, 21]. This platform empowers developers to craft specialized features tailored to their unique data lake requirements. Additionally, Delta Lake, developed by Databricks, represents another open-source data lake solution that aligns seamlessly with the Apache Spark APIs, enhancing the versatility and accessibility of data lakes within the ever-evolving data landscape.

2.2. Topic modeling

In the realm of computer science, Natural Language Processing (NLP) represents a challenging and vital field of study that empowers computers to comprehend human language within textual documents. Within this context, topic modeling approaches stand out as robust and intelligent algorithms frequently harnessed in NLP to uncover topics and extract valuable insights from unstructured text documents [22]. Broadly speaking, topic modeling techniques, particularly those grounded in LDA, discover many uses for text mining, social media analysis, information retrieval, and the broader domain of natural language processing. For example, the implementation of topic modeling based on LDA in the realm of social media analytics facilitates a deeper understanding of online community dynamics, enabling the extraction of valuable patterns not only from the content posted on social media platforms but also from the interactions and reactions of individuals in these virtual spaces. In a different research domain, topic modeling is employed in software architecture engineering, enabling the extraction of key topics from source code and facilitating the visualization of software similarities [23]. Figure 1 provides a visual representation of the diverse areas where these techniques find utilization. Thus, LDA provides a simple yet effective way to determine how each document is dispersed among different subjects and to measure the similarity between source inputs. Researchers have demonstrated the utility of this approach in tasks such as software refactoring and project organization. It is worth noting that various topic modeling techniques exist, with LDA standing out as one of the most renowned and widely adopted methods for this purpose as shown in Figure 1(b) [24]. The versatility of LDA can be used to a wide range of document types, such as social media posts, policy documents, collections of news items, political science texts, software engineering documents, medical literature, linguistics research, and even short-form content like tweets [7]. Our proposal builds upon recent advancements in topic modeling techniques, particularly focusing on the execution of LDA within the context of data lakes. While existing literature extensively covers various topic modeling methods, including LDA, their applicability to large-scale data lakes remains limited. Recognizing this gap, we conducted a thorough review of related studies and identified several key limitations inherent in current techniques when applied to the unique challenges posed by data lakes. These limitations include scalability issues, inadequate handling of heterogeneous data sources, and the lack of comprehensive solutions to prevent data swamps. By addressing these open challenges, our research seeks to provide practical solutions that enable effective topic modeling within the dynamic and complex environment of data lakes.

Moreover, while there are notable studies leveraging LDA for unstructured data analysis in related domains, there is a dearth of research specifically exploring the integration of LDA within the data lake framework. To illustrate, recent papers by Amara et al. [25], Kherwa and Bansal [26], and Sharif et al. [27] demonstrate the efficacy of LDA in extracting meaningful insights from unstructured data sources. However, these studies primarily focus on specific use cases and lack a holistic approach to data lake management. In contrast, our approach offers a comprehensive framework for topic modeling within data lakes, encompassing data ingestion, storage, and analysis. By comparing our methodology to existing approaches, we highlight the unique contributions of our research, particularly in addressing the scalability and heterogeneity challenges inherent in large-scale data lake environments. In summary, our work represents a significant advancement in the field of topic modeling, providing novel solutions to the open challenges associated with data lake management. By leveraging the power of LDA within the context of data lakes, we offer a practical and scalable approach to extracting actionable insights from heterogeneous data sources, thereby enhancing the value and usability of data lakes in modern data-driven enterprises.

Figure 1
Topic modeling techniques and their application areas

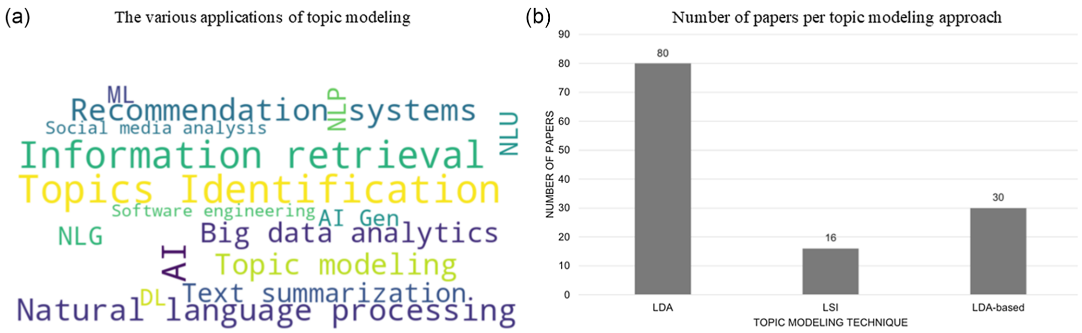


Table 1
Demonstrative accomplishments in topic modeling using LDA

Research suggestion	Goal of the research	Test dataset
Heintz et al. [28]	A method for uncovering linguistic patterns and a valuable asset for exploring conceptual metaphors	Wikidata
Lui et al. [29]	A method capable of identifying documents written in multiple languages	ALTW2010
Levy and Franklin [30], Zirn and Stuckenschmidt [31]	Examine political discord and its effects on the trucking industry	Online Regulatory Hub
Zhang et al. [32]	Exploring the distribution of user preferences	Yelp dataset
Zhang et al. [33]	Exploring geographic cluster discovery	Reuters-21587
Yang et al. [34]	Malicious Android app detection	Malicious application
Li et al. [35]	Examining the emotions analysis	Twitter dataset

Topic modeling techniques have demonstrated their substantial value and efficiency within the area of text analysis, enabling the extraction of latent patterns from documents and datasets using semantically driven analytics. This recognition drives our research to explore diverse topic modeling methodologies across various fields, encompassing aspects such as datasets, frameworks, resources, and real-world applications. As a result, topic modeling is expected to become more important in a variety of fields. In harmony complementing earlier research initiatives, we provide an extensive classification of the most recent topic modeling methodologies, particularly those rooted in the LDA model, covering a wide range of topics, such as dialectal science, software engineering, social media, politics, and others. As Table 1 illustrates, we also review significant researches within the field of computational linguistics that leverage topic modeling methodologies to enhance our understanding of this multifaceted domain.

3. Research Methodology

In this section, we illuminate upon the multifaceted procedures inherent in our proposal, providing insight into the sequential stages of our strategic implementation of the LDA model to effectively navigate and overcome the complexities associated with the data swamp challenge.

3.1. Materials and methods

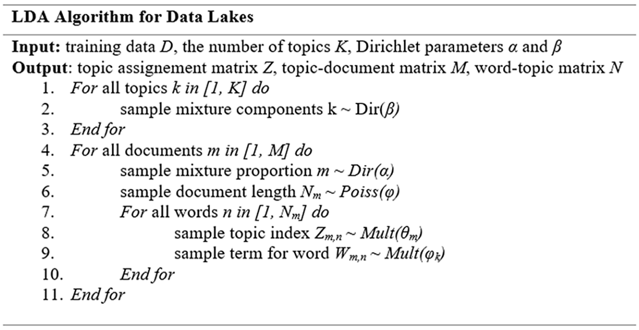
Efficiently ranking documents within the expansive search space of a data lake based on specific domain expertise is a substantial challenge with far-reaching implications. The task of identifying documents enriched with regular knowledge of a certain topic yet highly versatile operation. To address this challenge, we advocate for the adoption of a topic modeling strategy. Of the entire topic modeling techniques available, LDA stands out as a popular and well-respected option that has been used successfully in numerous text mining applications. Its role as a designated sub-task in information retrieval further underscores its significance and widespread adoption.

An essential aim of an expert discovery system revolves around evaluating the probability that a candidate, denoted as C, possesses expertise relevant to the input query Q. The system's approach involves assigning a ranking to these candidates, with each ranking based on the likelihood computation for the individual candidates within the search space. Consequently, the fundamental hurdle in the domain of expert discovery lies in the precise estimation of the probability $P(C|Q)$.

Topic modeling's aim is to produce a simplified depiction of the documents and the words. Every document is represented as a probabilistic mixture of subjects that have been taken out of a collection of documents. Additionally, topics are described as probability distributions across words in a document.

One of the currently favored approaches to topic modeling is LDA, initially presented by Blei et al. [36]. LDA acts as a generative model with probabilities. Its core premise lies in representing documents as random combinations of fundamental subjects, each represented by a word distribution. In practice, the words with the highest probabilities within these topics often offer valuable insights into their content and theme. LDA operates on the assumption that both topics and documents follow a Dirichlet prior distribution with consistency. Algorithm 1 illustrates the code skeleton for the determination of LDA specifications.

Algorithm 1. Pseudo-code for implementing the LDA algorithm in data lakes

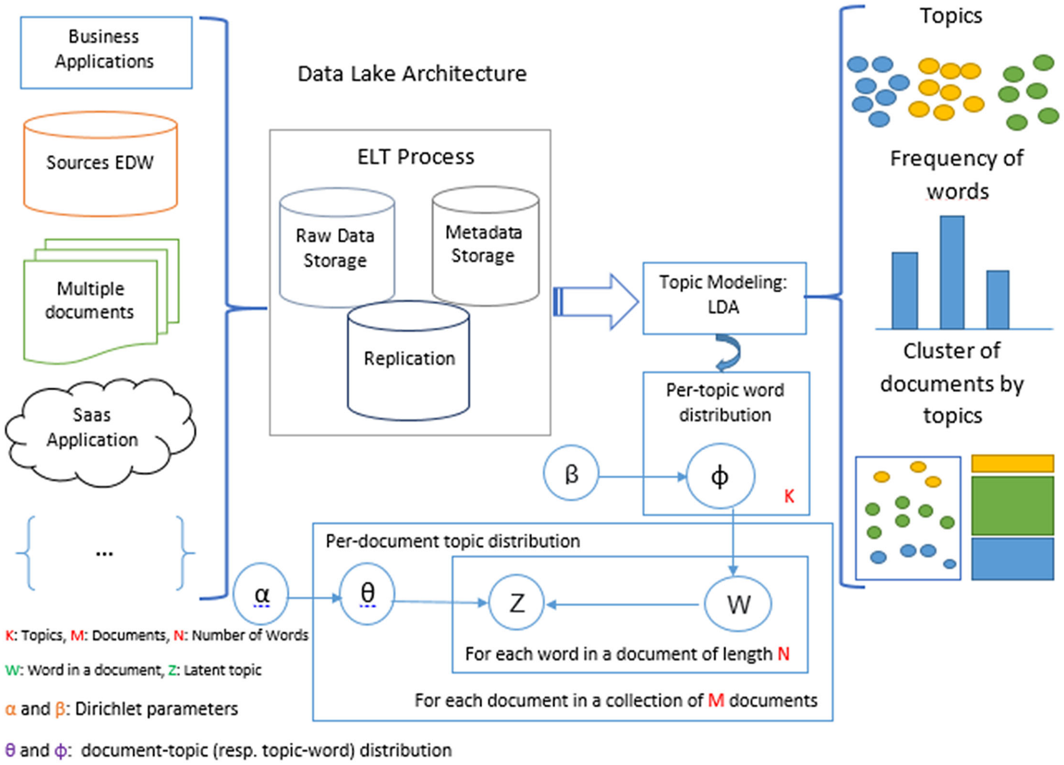


```
LDA Algorithm for Data Lakes
Input: training data $D$, the number of topics $K$, Dirichlet parameters $\alpha$ and $\beta$
Output: topic assignement matrix $Z$, topic-document matrix $M$, word-topic matrix $N$
For all topics $k$ in $[1, K]$ do
    sample mixture components k ~ Dir($\beta$)
End for
For all documents $m$ in $[1, M]$ do
    sample mixture proportion $m \sim Dir(\alpha)$
    sample document length $N_m \sim Poiss(\varphi)$
    For all words $n$ in $[1, N_m]$ do
        sample topic index $Z_{m,n} \sim Mult(\theta_m)$
        sample term for word $W_{m,n} \sim Mult(\varphi_k)$
    End for
End for
```

3.2. Proposed research framework design

To identify the key documents within the data lake, this study employs topic analysis as a data mining technique to explore its diverse resources. The extensive electronic and digital data archives offer significant opportunities and exert a substantial influence on knowledge discovery, information extraction, and analytical reasoning within the data lake environment. When it comes to handling extensive datasets, one prevalent method is topic modeling. Among the various techniques, LDA stands as a widely favored approach. LDA constructs topics by analyzing vocabulary from a document corpus. However, due to the inherent complexity of the data and the computational challenges posed by the analysis of massive document collections, often comprising millions of pages and billions of tokens, building meaningful topic models becomes a formidable task. Consequently, recent years have witnessed the emergence of several data processing frameworks, including Gensim, Mallet, and Spark, designed to address the intricacies of

Figure 2
The designed LDA approach for data lake systems


Business Applications
Sources EDW
Multiple documents
Saas Application
{ ... }
Data Lake Architecture
ELT Process
Raw Data Storage
Metadata Storage
Replication
Topic Modeling: LDA
Per-topic word distribution
β
ϕ
K
Per-document topic distribution
α
θ
Z
W
For each word in a document of length N
For each document in a collection of M documents
Topics
Frequency of words
Cluster of documents by topics
K: Topics, M: Documents, N: Number of Words
W: Word in a document, Z: Latent topic
α and β: Dirichlet parameters
θ and ϕ: document-topic (resp. topic-word) distribution

examining vast volumes of unlabeled data across diverse domains in a scalable and efficient manner. This paper presents a working prototype of the LDA approach applied with the Gensim library to extract topics that encapsulate the rich diversity of data stored within data lakes. Figure 2 presents a thorough synopsis of the suggested framework for this study.

3.3. Assessment metrics

Topic models that are probabilistic, like LDA, are widely favored methods for text analysis due to their ability to provide both predictive and latent topic representations within a corpus. While it is frequently difficult to assess these assumptions due to their unsupervised learning processes, there has been a longstanding belief in the general relevance and value of the latent space they uncover. However, it is crucial to establish a means of objectively evaluating several topic models to ascertain if a trained model is demonstrably successful or failed. Achieving a fair assessment of model caliber is imperative, and the ideal approach involves one common universal and optimizable metric. Several evaluation approaches are commonly employed for models like LDA, encompassing visual inspections, containing the top N words, subjects, and documents; vital assessment metrics that reflect topic coherence and model semantics; individual judgment to assess the degree to which the topics that were taken out correspond with the document's content; and third-party criteria for assessment that assess how well the model performs in completing assigned duties. In this paper, we delve deeper into the concept of topic coherence, a crucial assessment measure, and explore its application as a statistical support tool for making informed model choices.

Before delving into the concept of topic coherence, let us examine the perplexity metric simply. Among the intrinsic evaluation metrics, perplexity is a frequently used statistic for assessing language models. It measures a model's degree of unexpectedness when encountering new, never-before-seen data and is represented as the held-out test set's standardized log-likelihood. By concentrating on the log-likelihood measure, confusion can be viewed as a way to evaluate the possibility that the model will forecast unexpected data based on its previous training. However, the relationship between human judgment and prediction likelihood (or ambiguity) is often lacking and occasionally shows some degree of negative association, according to current studies. Consequently, optimizing for perplexity may lead to outcomes that do not align with human assessment. This limitation of perplexity measurement prompted further investigation aimed at replicating human judgment, giving rise to the concept of topic consistency.

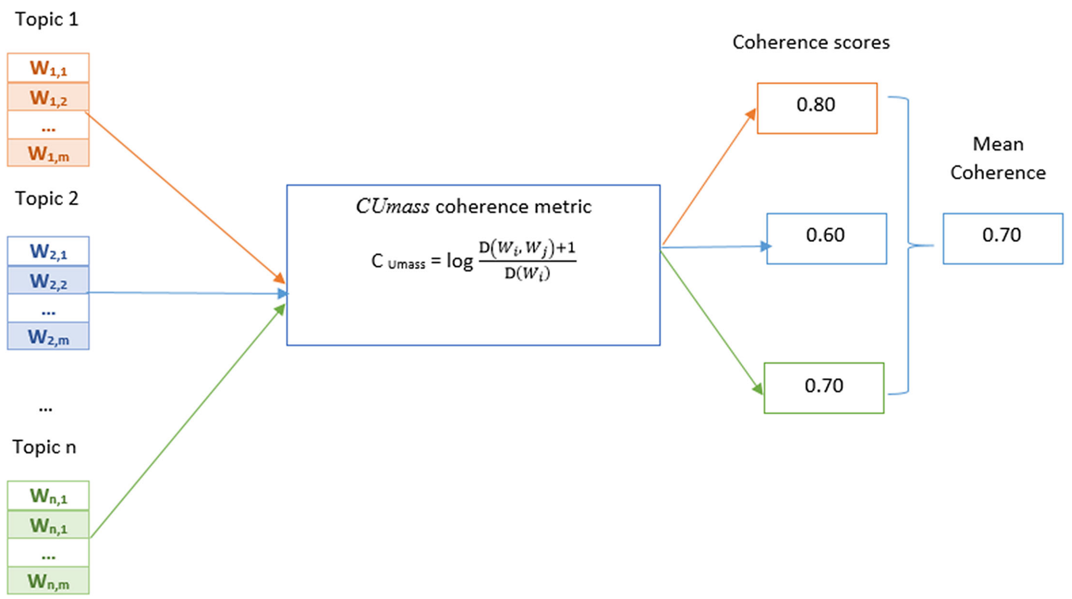
The concept of topic coherence incorporates multiple metrics into an all-encompassing structure intended to evaluate the reasonable coherence between topics that a model has identified. Among frequently employed consistency metrics, C_{Umass} coherence stands out. This metric constructs topic representations by analyzing their co-occurrences and subsequently computes the score utilizing Normalized Pointwise Mutual Information and cosine similarity. An example of the topic coherence assessment procedure is shown in Figure 3.

Where $D(w_i, w_j)$ denotes the co-occurrence frequency of terms w_i and w_j within documents, and $D(w_i)$ represents the standalone frequency of word w_i. A higher coherence score is indicative of better performance.

4. Results and Discussions

In this section, we highlight the key discoveries from our proposed study, commencing with a comprehensive exposition of

Figure 3
Evaluating coherence scores across diverse topics



the experimental configuration and proceeding to an exhaustive exploration of the result analysis.

4.1. Experimental setup

This study aims to uncover the latest topics within data lake resources, while filtering out irrelevant data by pinpointing significant issues like data swamps. To achieve this, we collected data spanning for four years, from 2018 to 2022, from reputable scientific databases, including Scopus, Thomson Reuters ISI, Springer, Elsevier, Web of Science, PubMed, USPTO, and others. The corpus was made up of publicly accessible documents that were taken from reliable research articles in the data lake industry. Certain criteria were used in the selection process: the material had to be globally peer-reviewed, published by a respectable publisher, and indexed. Further, the hardware and software specifications of the experimental device are presented in Table 2.

Table 2
Technical specifications of the experimental device

Hardware configuration	Software configuration
CPU: 1.80GHz Intel(R) Core(TM) i7-10510U	Windows 10 is the operating system
RAM: 16GB	Using the Gensim framework
DISK: 512 SSD, 1TO HDD	Python is the programming language.
GPU: GeForce MX250 from NVIDIA	Storage area: Repository of data lakes

In preparation for topic modeling using LDA, we conducted rigorous preprocessing steps to ascertain the superiority and suitability of the input data. This involved several key processes, including data cleaning, tokenization, and document parsing. Specifically, we employed advanced techniques for tokenization, breaking down each document into individual tokens or words, while removing irrelevant characters, punctuation marks, and stop words. Additionally, we implemented stemming and lemmatization algorithms to normalize the text, reducing inflectional forms and variants to their base or root forms. Furthermore, we applied domain-specific filtering to eliminate noise and irrelevant terms, ensuring that only relevant and meaningful words were retained for subsequent analysis. This preprocessing pipeline was meticulously designed to optimize the input data for LDA, facilitating more accurate and insightful topic modeling results. These preprocessing steps were primarily implemented using powerful Python libraries such as scikit-learn, NLTK, and spaCy, which provided robust and efficient tools for text processing and analysis.

4.2. Discussion of the results

Finding the latent issues that are hidden inside the data lake is the main goal of this study. To accomplish this, we employ semantic analysis to select the best 10 keywords from a list of 498 that were included in the abstracts of 15 research papers. However, finding the right range of topics can be very difficult, especially if the data were unknown beforehand. Yet, we propose to use topic coherence to determine the optimal number of topics given two preset LDA hyper-parameters ($\alpha = 0.1$ and $\beta = 0.89$). Then, the optimal hyper-parameter combination for the topic model is determined through the utilization of the hyper-parameter grid search technique. This method systematically explores a predefined hyper-parameter space by evaluating various combinations of α and β values. To evaluate each model's performance, we rely on perplexity and log-likelihood measures. This evaluative metric aids in achieving the highest level of accuracy while concurrently reducing processing time. The optimal number of topics is depicted in Figure 4.

Conversely, the success of LDA critically depends on determining the optimal values for its hyper-parameters, striking a balance for efficient topic processing with sustained high coherence, as evidenced in Table 3. Although various innovative methods exist for hyper-parameter selection, this research paper

Figure 4
Determining the optimal number of LDA topics

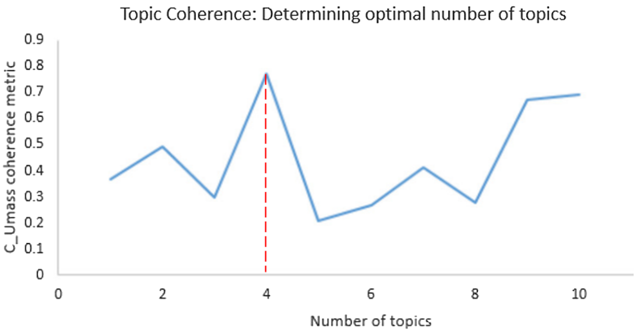


Table 3
Suitable LDA tuning parameters

α	β	C_{Umass}
0.10	0.89	0.672
0.13	0.90	0.657
0.72	0.92	0.628
0.82	0.93	0.597

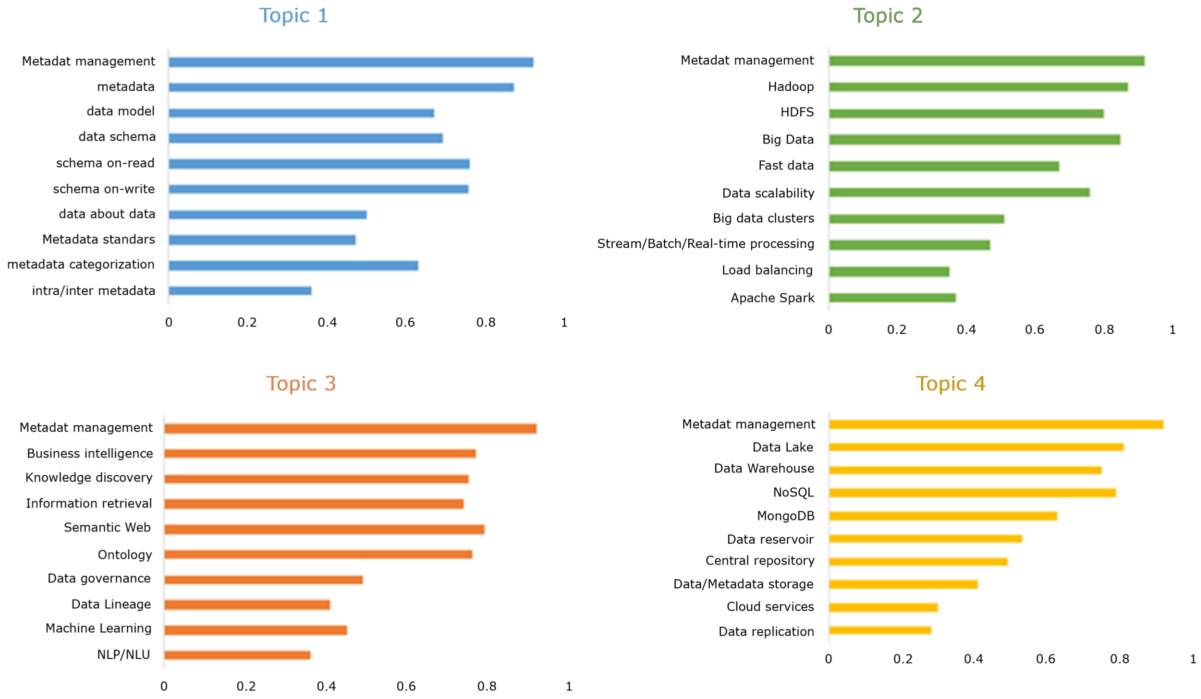
adheres to values that resulted in the highest C_{Umass} score for $K = 4$, aligning with the findings from our comprehensive analysis.

In parallel, a bar chart, featured in Figure 5, is utilized to visually represent word frequencies, enhancing the intuitive grasp of the data analysis findings. This visualization not only provides a detailed exploration of linguistic patterns but also aligns with the broader context of topic modeling applied among the dynamic environment of a data lake system. By visually dissecting word frequencies, our approach illuminates the latent topics prevalent in the dataset, offering a nuanced understanding of semantic structures and contributing to the effective management and interpretation of vast and diverse data within the data lake ecosystem.

The findings show that "metadata management" is the significant "top ten" keyword in the data lake system, compared to "data catalog," "cloud services," "ontology," "NoSQL databases," "Semantic Web," "machine learning models," "apache hadoop ecosystem," "data-driven business intelligence," "management-centric data governance," and "big data architectures." Hence, the utilization of the LDA model proves to be well-suited for the intention of data analysis and the encapsulation of hidden subjects, providing us with an open understanding of the lakes' contents and, as a result, preventing problems associated with data lakes, such the notorious "data swamp."

In our comprehensive analysis of the results, we not only highlight the strengths of our approach but also critically examine its limitations and areas for improvement. Firstly, we acknowledge that while our approach demonstrates robust performance across various languages, there are inherent limitations in the languages supported due to the availability and quality of textual data in different languages. Future efforts could focus on expanding language support to ensure broader applicability and inclusivity. Additionally, we recognize the computational complexity associated with larger datasets, which may pose challenges in terms of processing time and resource requirements. To address this, we emphasize the importance of optimizing algorithmic efficiency and exploring parallel computing strategies to enhance scalability. Furthermore, in our analysis of model predictions, we meticulously scrutinize errors and weaknesses, particularly in the identification of specific topics. Through a thorough examination of misclassified topics and ambiguous predictions, we gain valuable insights into the underlying factors contributing to model performance limitations. By elucidating these nuances, we pave the way for future research endeavors aimed at

Figure 5
Word distribution among topics in data lake documents



refining the model architecture, enhancing feature representation, and incorporating domain-specific knowledge to improve topic distinction accuracy. Overall, our discussion not only underscores the achievements of our approach but also serves as a catalyst for ongoing refinement and innovation in the field of topic modeling.

5. Conclusion and Future Perspectives

In this paper, we have extensively delved into the crucial role that LDA plays in modeling data lakes and categorizing their contents into four distinct categories, effectively addressing and mitigating the prevalent challenge of data swamps. Our exploration has not only shed light on the significance of LDA-generated topics as reference points but has also highlighted their utility as valuable tools for authors navigating the vast landscape of documents within the data lake. Our investigation covered a broad spectrum of aspects, encompassing sophisticated topic extraction techniques, rigorous performance assessments, and meticulous inference parameter estimation. Through the implementation of the LDA approach as a topic modeling algorithm, our study has consistently demonstrated superior performance, particularly evident in the achieved high levels of topic coherence. The critical novelty and contributions that our work has made in the innovative application of LDA to tackle the challenges of data lakes, offering a robust solution for categorization and management.

Looking forward, our findings open up a promising future perspective—the integration of labeled documents. This forward-looking trajectory involves associating informative labels with the identified topics. By harnessing the implicit information embedded in document contexts, this approach aims to significantly enhance predictive capabilities. The integration of labeled documents represents a substantial leap towards achieving more precise predictions and gaining deeper insights, thereby charting an exciting course for future research endeavors. Additionally, we propose practical applications for our approach in various domains, including information retrieval, recommendation systems, and decision support systems, where accurate categorization of data lake contents is essential for informed decision-making. Furthermore, as part of our ongoing research, we anticipate exploring and implementing various advanced topic modeling techniques beyond LDA. This broader exploration aims to enrich the spectrum of methodologies available for uncovering latent patterns in data lakes, including worthwhile extensions like sentiment analysis of extracted topics. By continuously refining and expanding our methodologies, we aim to promote the development of data lake management techniques and facilitate deeper insights into complex data ecosystems.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Mohamed Cherradi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Anass El Haddadi:** Supervision.

References

[1] Alnoukari, M. (2021). From business intelligence to big data: The power of analytics. In A. Azevedo & M. Santos (Eds.), *Integration challenges for analytics, business intelligence, and data mining* (pp. 44–62). IGI Global. https://doi.org/10.4018/978-1-7998-5781-5.ch003

[2] Ruzgas, T., & Dabulytė-Bagdonavičienė, J. (2017). Business intelligence for big data analytics. *International Journal of Computer Applications Technology and Research*, *6*(1), 1–8. https://doi.org/10.7753/IJCATR0601.1001

[3] Huang, S. C., McIntosh, S., Sobolevsky, S., & Hung, P. (2017). Big data analytics and business intelligence in industry. *Information Systems Frontiers*, *19*, 1229–1232. https://doi.org/10.1007/s10796-017-9804-9

[4] Olawoyin, A. M., Leung, C. K., & Cuzzocrea, A. (2021). Open data lake to support machine learning on arctic big data. In *2021 IEEE International Conference on Big Data*, 5215–5224. https://doi.org/10.1109/BigData52589.2021.9671453

[5] Cherradi, M., & El Haddadi, A. (2023). DLDB-service: An extensible data lake system. In *Emerging Trends in Intelligent Systems & Network Security: Conference Proceedings*, 211–220. https://doi.org/10.1007/978-3-031-15191-0_20

[6] Rajapaksha, M., & Silva, T. (2019). Semantic information retrieval based on topic modeling and community interests mining. In *2019 Moratuwa Engineering Research Conference*, 60–65. https://doi.org/10.1109/MERCon.2019.8818935

[7] Maini, E., Venkateswarlu, B., & Gupta, A. (2018). Data lake – An optimum solution for storage and analytics of big data in cardiovascular disease prediction system. *International Journal of Computational Engineering & Management*, *21*(6), 33–39.

[8] Khine, P. P., & Wang, Z. S. (2018). Data lake: A new ideology in big data era. In *4th Annual International Conference on Wireless Communication and Sensor Network*, 17, 03025. https://doi.org/10.1051/itmconf/20181703025

[9] Terrizzano, I., Schwarz, P., Roth, M., & Colino, J. E. (2015). Data wrangling: The challenging journey from the wild to the lake. In *Conference on Innovative Data Systems Research*.

[10] Marty, R. (2015). *The security data lake: Leveraging big data technologies to build a common data repository for security*. USA: O'Reilly.

[11] Sitarska-Buba, M., & Zygala, R. (2020). Data lake: Strategic challenges for small and medium sized enterprises. In M. Hernes, A. Rot & D. Jelonek (Eds.), *Towards industry 4.0—Current challenges in information systems* (pp. 183–200). Springer.

[12] Walker, C., & Alrehamy, H. (2015). *Personal data lake with data gravity pull*. In *2015 IEEE Fifth International Conference on Big Data and Cloud Computing,* 160–167. https://doi.org/10.1109/BDCloud.2015.62

[13] Fang, H. (2015). Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems*, 820–824. https://doi.org/10.1109/CYBER.2015.7288049

[14] Ben, S. (2018). *Architecting data lakes*. USA: O'Reilly.

[15] Cherradi, M., Bouhafer, F., & El Haddadi, A. (2023). Data lake governance using IBM-Watson knowledge catalog. *Scientific African*, *21*, e01854. https://doi.org/10.1016/j.sciaf.2023.e01854

[16] Inmon, B. (2016). *Data lake architecture: Designing the data lake and avoiding the garbage dump*. USA: Technics Publications.

[17] Madera, C., & Laurent, A. (2016). The next information architecture evolution: The data lake wave. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, 174–180. https://doi.org/10.1145/3012071.3012077

[18] Ravat, F., & Zhao, Y. (2019). Data lakes: Trends and perspectives. *Database and Expert Systems Applications: 30th International Conference*, 304–313. https://doi.org/10.1007/978-3-030-27615-7_23

[19] Mathis, C. (2017). Data lakes. *Datenbank-Spektrum*, *17*(3), 289–293. https://doi.org/10.1007/s13222-017-0272-7

[20] Suriarachchi, I., & Plale, B. (2016). Crossing analytics systems: A case for integrated provenance in data lakes. In *2016 IEEE 12th International Conference on e-Science*, 349–354. https://doi.org/10.1109/eScience.2016.7870919

[21] Cherradi, M., & El Haddadi, A. (2022). *Data lakes: A survey paper*. In *Proceedings of the 6th International Conference on Smart City Applications*, 823–835. https://doi.org/10.1007/978-3-030-94191-8_66

[22] Kim, M., & Kim, D. (2022). A suggestion on the LDA-based topic modeling technique based on ElasticSearch for indexing academic research results. *Applied Sciences*, *12*(6), 3118. https://doi.org/10.3390/app12063118

[23] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

[24] Costa Silva, C., Galster, M., & Gilson, F. (2021). Topic modeling in software engineering research. *Empirical Software Engineering*, *26*(6), 120. https://doi.org/10.1007/s10664-021-10026-0

[25] Amara, A., Hadj Taieb, M. A., & Ben Aouicha, M. (2021). Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis. *Applied Intelligence*, *51*, 3052–3073. https://doi.org/10.1007/s10489-020-02033-3

[26] Kherwa, P., & Bansal, P. (2019). Topic modeling: A comprehensive review. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(24), e2. https://doi.org/10.4108/eai.13-7-2018.159623

[27] Sharif, M., Maskat, R., Baharum, Z., & Maskat, K. (2023). A scoping review of topic modelling on online data. *Indonesian Journal of Electrical Engineering and Computer Science*, *31*(3), 1633–1641.

[28] Heintz, I., Gabbard, R., Srivastava, M., Barner, D., Black, D., Friedman, M., & Weischedel, R. (2013). Automatic extraction of linguistic metaphors with LDA topic modeling. In *Proceedings of the First Workshop on Metaphor in NLP,* 58–66.

[29] Lui, M., Lau, J. H., & Baldwin, T. (2014). Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, *2*, 27–40. https://doi.org/10.1162/tacl_a_00163

[30] Levy, K., & Franklin, M. (2014). Driving regulation: Using topic models to examine political contention in the US trucking industry. *Social Science Computer Review*, *32*(2), 182–194. https://doi.org/10.1177/0894439313506847

[31] Zirn, C., & Stuckenschmidt, H. (2014). Multidimensional topic analysis in political texts. *Data & Knowledge Engineering*, *90*, 38–53. https://doi.org/10.1016/j.datak.2013.07.003

[32] Zhang, Y., Chen, M., Huang, D., Wu, D., & Li, Y. (2017). iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems*, *66*, 30–35. https://doi.org/10.1016/j.future.2015.12.001

[33] Zhang, L., Sun, X., & Zhuge, H. (2015). Topic discovery of clusters from documents with geographical location. *Concurrency and Computation: Practice and Experience*, *27*(15), 4015–4038. https://doi.org/10.1002/cpe.3474

[34] Yang, X., Lo, D., Li, L., Xia, X., Bissyandé, T. F., & Klein, J. (2017). Characterizing malicious Android apps by mining topic-specific data flow signatures. *Information and Software Technology*, *90*, 27–39. https://doi.org/10.1016/j.infsof.2017.04.007

[35] Li, Y., Zhou, X., Sun, Y., & Zhang, H. (2016). Design and implementation of Weibo sentiment analysis based on LDA and dependency parsing. *China Communications*, *13*(11), 91–105. https://doi.org/10.1109/CC.2016.7781721

[36] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

How to Cite: Cherradi, M., & El Haddadi, A. (2025). Enhancing Data Lake Management Systems with LDA Approach. *Journal of Data Science and Intelligent Systems*, *3*(1), 58–66. https://doi.org/10.47852/bonviewJDSIS42022312