International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Observations and Expectations on Recent Developments of Data lakes

Fasi Ahmed Parvez Mohammad[1*], Dr. Awakash Mishra[2]

Submitted: 03/05/2024 **Revised**: 16/06/2024 **Accepted**: 23/06/2024

Abstract: The idea of a data lake was first conceived more than a decade ago, and despite the fact that considerable advancements have been achieved in research as well as applications, there are still a great deal of problems and difficulties that have not been overcome. This paper provides a comprehensive overview of the field's recent progress, including our findings and predictions for the future of both theory and practice. First, we will examine the common terms used to describe data lakes. In the following steps, we will conduct an in-depth investigation into a variety of aspects of data lakes. This will include conducting a comprehensive review of metadata and the challenges that are associated with it, delving into the nuanced complexities of the information contained within data lakes, and analyzing distinctive characteristics such as the mixed lazy and eager approach that is utilized throughout the data lake lifecycle. Furthermore, we analyses the dynamics of the link between data lakes and data mining, look into how data lakes interact with data warehousing methods, and shed light on the latest developments in data lake applications. Throughout these analyses, we share our observations, culminating in the paper's conclusion, where we express our aspiration for the development of a unified framework to evaluate data lake lifecycles.

Keywords: *data lake, metadata management, data profiling, bigdata analytics, data ware house.*

1. Introduction

The inception of the data lake concept dates back to 2010, as proposed by James Dixon [17]. Dixon likened the data lake to "a large body of water in a more natural state," distinguishing it from the structured nature of a data mart. Unlike a data mart, which serves as a purified and organized store of bottled water for easy consumption, the data lake operates as an open reservoir continuously receiving content from diverse sources. Users are analogous to visitors at a lake, able to explore, immerse themselves, or extract samples [17]. This data lake phenomenon mirrors a paradigmatic shift, reminiscent of the evolution in the database management systems field. While the 1980s saw object-oriented databases reinforcing rigid schemas, subsequent developments, including XML and contemporary big data trends [43], have embraced an alternative direction, liberating data storage and manipulation from the confines of rigid schemas. In contrast to data warehousing [9], which adheres to strict schemas for managing consolidated data, data lakes adopt an inclusive stance. They welcome diverse data types and employ them in their raw state, irrespective of whether the data is structured, semi-structured, or unstructured.

Illustrating the versatility of the data lake concept, one notable example is found in the realm of COVID-19 research and development. A public data lake is accessible at https://aws.amazon.com/covid-19-data-lake, showcasing the application of data lakes in providing a dynamic and inclusive environment for diverse data types.This evolution from rigid schemas to more flexible and encompassing data storage mechanisms signifies a departure from traditional database management approaches. The rebellious direction embraced by data lakes aligns with the broader trends observed in the evolving landscape of data management.

Over time, there has been a growing interest, or curiosity, around data lakes. However, the landscape of online materials related to data lakes is characterized by diversity, with significant variations in quality. It is essential to underscore that, in contrast to its "predecessor," the data warehouse, there remains a notable absence of a systematic framework for delving into data lake research. In this area, researchers have expressed concerns and pointed out gaps that need to be filled (e.g., [20]). Many people who are new to data lakes ask how to properly manage the massive "lake" of literature on the topic, especially those who are already well-versed in database-related IT but are just starting to delve into the details. On the other hand, how can one quickly understand the main ways to approach this important subject? This investigation becomes more complicated due to the lack of a formal study framework for data lakes. Researchers and practitioners may find it difficult to sift through the mountain of information available online. For individuals accustomed to more organised methods, such as data warehouses, the necessity for a consistent basis to understand data lakes

[1*] *Ph.D. Research Scholar, Department of Computer Science, Maharishi School of Engg. & Tech., MUIT University, Lucknow, U.P. E-Mail: parvez40509@gmail.com,*
[2] *Professor, Department of Computer Science, Maharishi School of Engg. & Tech., MUIT University, Lucknow, U.P.*
***Corresponding Author:** *Mr. Fasi Ahmed Parvez Mohammad*
**E-Mail: parvez40509@gmail.com,*

becomes apparent, resolving worries and doubts. The pursuit of a simplified and easily understandable knowledge becomes paramount in this dynamic landscape as the importance of data lakes grows. People want to not only survive in the vast data lake domain, but thrive in it, thus there's an obvious need for abundant resources and organised direction.

This paper aims to provide a partial response to the growing interest in data lakes from a learner's standpoint. Recognizing the author's limitations in delivering a complete review, the objective is to share observations and insights derived from an incomplete survey. The goal is to lay the groundwork for future comprehensive research that meet the rising standards in this field. The subsequent sections of this article provide a partial literature review of resources online that are devoted to data lakes. Although the quality of these resources varies, a good place to start is with tutorials and surveys put out by respected organisations and sources; from there, you can build your research by consulting publications that have been heavily cited. Data warehousing approaches [24,25] and database management systems (DBMS) [42] are foundational to understanding data lakes and how to navigate them. References [40, 28] have been selected as possible entryways into the vast terrain of data lake literature, and specific papers have been recognised as appropriate starting places. The data lake landscape is critically examined from a high-level perspective in [20]. In addition, reference [33] focuses on data lake potential and difficulties, including data input, extraction, cleansing, discovery, metadata management, integration, and versioning; it is mainly a lesson with just an expanded abstract online.

A more comprehensive and detailed examination of data lakes may be found in [23]. Although learners may find it difficult to understand all of the material, the paper's format gives a good summary of the most important aspects of data lakes. The structure includes a data lake idea definition, a data lake historical overview, and the authors' data lake architectural introduction. Also included are certain criteria for categorizing data lake solutions built on top of this architecture. Here are the three distinct layers that make up a data lake: the ingestion layer is in charge of metadata modelling and extraction; the maintenance layer is in charge of organizing and preparing datasets; data interaction and quality improvement; schema evolution; and lastly, the exploration layer is in charge of query-driven data discovery and querying of heterogeneous data. For a more specific aspect of data lake management, [48] is recommended for its insights into data lake ingestion management. As some terms used in this overview will be explained in later sections, the groundwork is laid for a more comprehensive understanding of the intricacies of data lakes in the subsequent parts of this paper.

In the subsequent sections of this paper, our focus shifts towards providing observations on recent developments in data lakes, with a particular emphasis on catering to a learner's perspective. We delve into various aspects related to data lakes, identifying notable issues, some explicitly stated in the literature, and others implied. The following organization outlines the key areas of examination. In Section 2, we delve into the specifics of general terminology associated with data lakes. Moving forward, Section 3 conducts an examination of metadata information, while Section 4 provides a detailed description of the granules involved in data lakes. Section 5 delves into the intricacies of the mixed lazy and eager approach adopted in the lifecycle of data lakes, and Section 6 explores the extraction of knowledge patterns from these expansive repositories. In Section 7, we explore the nature of the connection between data lakes and data warehouses, explaining how they work together and how they might complement one other. Section 8 offers brief remarks on several data lake uses, moving us towards practical applications. Section 9 concludes by analysing the viewpoints of experts regarding the inherent difficulties of data lakes. In addition, we sketch out our plans for further studies in this area, stressing the apparent necessity of a complete framework to evaluate the data lake lifetime. This structured approach guarantees a methodical examination of important aspects, providing learners with a comprehensive understanding of the dynamic data lake landscape.

2. Terminology

The structural nuances of data reservoirs must, nevertheless, be comprehended. The architecture of a data lake delineates the system's framework and components, providing essential guidance on how data is stored, organized, and harnessed. While on-premise data lakes continue to be widely adopted, the rise of cloud computing has ushered in cloud data lakes as a compelling alternative, offering elastic compute services for on-demand data analysis stored in the cloud. Notably, the Lambda architecture [39] has gained prominence as a widely used data lake architecture in industrial applications. Within academic research, diverse perspectives on data lake architecture have been explored, as evidenced in discussions found in research papers like [21]. Although a unanimous consensus on the precise structure of a data lake architecture is lacking, most proposals involve multiple layers, encompassing data ingestion, maintenance, and exploration [23].

Another pivotal term in the data lakes landscape is the concept of modeling. With the absence of predefined schemas in data lakes, questions may arise about the relevance of modeling. Even when there is no universally agreed-upon format for storing data, the necessity for data models and modeling can still arise. For instance, when contemplating data storage, it becomes crucial to consider the underlying Hadoop paradigm. Metadata modeling, extensively covered in data lake literature, constitutes a vast and intricate area that warrants further research. To navigate the intricate realm of data lakes, familiarity with data lake design, modeling, and other technical terminologies is indispensable. At first glance, data lakes may appear as vast reservoirs of raw data lacking any structure. However, a deeper exploration reveals that robust architecture and modeling are imperative for effectively managing and leveraging the diverse data sources that constitute this dynamic ecosystem.

Additional terminology has been developed as a result of the distinctive characteristics of the data lake lifecycle. As an example, consider the original presentation of the concept of query discovery in [31]. When searching for a query or transformation that changes the form of data, this phrase describes the process of finding the correct operators to join, nest, group, connect, and transform data in order to achieve a desired format. Section 5 will bring us back to the idea of query-driven data discovery, but in the modern data science landscape, data analysis necessitates discovering new data that seamlessly integrates, combines, or aggregates with existing data.

3. Metadata, schema matching and ontology

Recent developments in data lakes underscore the critical roles of metadata, schema matching, and ontology in reshaping the landscape of data management and analytics. Metadata, serving as the informational backbone, offers insights into data sources, lineage, quality, and usage. Automated metadata management tools are gaining prominence, enhancing data governance and efficiency. Schema matching techniques, leveraging advancements in machine learning, address the challenge of diverse data formats, ensuring more scalable and accurate alignment of disparate schemas. Simultaneously, ontologies, representing knowledge and relationships, play a vital role in semantic data modeling, fostering a shared understanding of data elements. The integration of ontological models enhances search capabilities within data lakes, enabling more intelligent and context-aware queries. Together, these advancements contribute to a more intelligent, automated, and interoperable data management environment, empowering organizations to efficiently harness insights from the vast and heterogeneous data reservoirs within modern data lakes.

Metadata is vital in data lakes for describing and classifying the large datasets stored there as data comes from so many different places. Metadata is the main entry point for interacting with any dataset in a data lake; it is the bedrock upon which understanding and altering the stored data rests. This section gives a concise introduction, bringing insights into the topic of metadata extraction, while metadata in data lakes has been thoroughly covered in many research articles. Metadata extraction strategies are demonstrated in [29], which provides an example of how to extract metadata from datasets stored in a data lake. Metadata plays a more crucial role in dataset discovery, also known as data profiling, to satisfy customers' information demands [1,34]. Metadata ontologies are essential for managing metadata semantics. They contribute to the improvement of dataset discovery.

To delve further into the topic, see [30] for an analysis of data profiling in a broader perspective. Author in [3] presents a prototype system for information profiling that uses ontologies to improve metadata interpretation in the real world. The system includes an ontology alignment component. The various facets of metadata utilization in data lake scenarios are further highlighted in [2], which proposes a proximity mining strategy for pre-filtering schema matching. It must be emphasized that the following section will elaborate on the subject of information granularity, which is closely related to the metadata examination. The significance of metadata in guiding our comprehension and engagement with the diverse and detailed material stored in data lakes is highlighted by this correlation.

4. Granularities in data lakes

In order to manage objects and data in a data lake, it is necessary to handle metadata at various levels of granularity, which entails using similarity metrics to build these granules. Here, the metadata approach proposed in [18] is illustrative; it permits information capture at different granularities, allowing classification that incorporates both specifically defined data components and broader data ranges. When it comes to granularities, an issue that impacts many sections of data lakes, entity resolution (ER) stands out. Data integration isn't complete without entity resolution (ER), which focuses on bringing together different sources' representations of entities. Methods like blocking and meta-blocking are used to manipulate information granules. In [4], the idea of query-driven entity resolution is presented for use in data lakes. Blocking is used to limit comparisons to similar entities by grouping them into entity blocks, and meta-blocking is used to reorganize a block in order to reduce repetition. This methodology results in the formulation of an entity resolution-enriched query plan, processed using a newly proposed block join operator. The examination of such information granules and the management of associated uncertainties fall under the purview of granular computing (GrC) [46]. Employing rough set theory, a foundational aspect of GrC, [44] integrates data silos from diverse organizations within a data lake, optimizing operational business processes and enhancing data quality and efficiency.Despite the potential of granular computing to significantly contribute to metadata management in data lakes, the current landscape does not align with these expectations. A search for "GrC for data lakes" often yields results related to governance (G), risk (R), and compliance (C), commonly known as GRC in data lakes literature [5], rather than granular computing. The expectation is that research endeavors focusing on data lakes from the granular computing perspective will gain traction, reshaping search results in the years to come.

5. Mixedlazy and eager approach in data lake life cycle

Recent developments in data lakes have witnessed a nuanced approach to their lifecycle, blending elements of both laziness and eagerness. The lazy approach emphasizes deferred execution and on-demand processing, allowing for the efficient storage of raw, unprocessed data until it's needed for analysis. This minimizes upfront processing costs and accelerates data ingestion. On the other hand, the eager approach involves preprocessing and indexing data upon ingestion, enabling faster query performance but requiring more upfront computational resources. This mixed approach seeks to strike a balance, leveraging lazy loading for cost-effective storage and eager processing for quicker analytical insights when needed. As data lakes evolve, this blended strategy anticipates increased flexibility and adaptability in managing the diverse and dynamic nature of data, catering to the specific needs of both storage efficiency and real-time analytics within the data lake ecosystem.

The philosophy underpinning the data lake concept centers around acquiring data in its raw form before processing or transforming it. A lack of rigorous methodology is implied by this operational approach in data lakes. There is a significant difference, though, when contrasted with data warehouses, where analysis is typically performed after data consolidation. When dealing with data lakes, it's common to need a data analysis method to unearth the necessary information, which indicates a proactive strategy. Although it's not a problem in and of itself, this dynamic has far-reaching consequences for the data lake lifespan as a whole. A more thorough analysis sheds light on this matter. Data lakes deviate from the conventional Extract, Transform, and Load (ETL) methodology used by data warehouses, as highlighted in [28]. Alternatively, data lakes adhere to the ELT paradigm, which means they keep data unprocessed until it's needed for queries or applications. This is in sharp contrast to data warehouses, which store aggregated historical data and are hence the backbone of efficient data mining and analysis.

Although data lakes and data warehouses both store data, there is a dearth of conversation about accessing and mining the full lake. In [23], the authors explain this occurrence by outlining a two-pronged strategy for exploring data lakes: finding lakes with strong dataset relationships, and then giving users a single interface to query all their disparate data sets. As a result, querying and analysis are not the final stages of the data lake lifecycle; instead, they are intertwined with activities across various layers. In the context of data lakes, the focus on data mining and machine learning may not revolve solely around discovering hidden knowledge patterns stored in the entire data lake. Rather, the emphasis is often on tasks related to identifying relevant datasets aligned with users' interests. This paradigm shift underscores the emergence of query-driven data discovery (Section 2), with research fervor evident in works like [7,14,15,47,48]. This evolving perspective reflects the dynamic nature of data lake exploration, illustrating ongoing efforts to tackle challenges and enhance the efficacy of querying and discovering meaningful insights within the extensive and ever-changing data lake environment.

6. Knowledge pattern extraction and knowledge lakes

Expanding upon the preceding discussion, it's crucial to underscore that, unlike data warehouses, data mining within the framework of data lakes doesn't require waiting for the entire data lake to be fully constructed. Although this difference has been made, there has been comprehensive research on data lakes, as seen in publication [10]. In order to improve management efficiency, this research uses a network-based method to identify visual knowledge patterns in a data lake. In [6], the notion of an intelligent knowledge lake is introduced, which is another significant improvement. The authors' earlier proposal of a knowledge lake—a contextualized data lake with related algorithms—forms the basis of this idea. Helping AI and data analytics work together more seamlessly is the goal. The intelligent knowledge lake is designed to facilitate the learning of AI applications using data that has been contextualized. Because of this, cognitive support may be created and business operations can be automated. This, in turn, makes knowledge-intensive activities easier or even generates new rules for business analytics in the future. This innovative approach signifies a convergence of AI and data analytics within the dynamic environment of a data lake, promising transformative potential for businesses seeking advanced insights and automation capabilities.

7. Data warehouses and data lakes

A significant amount of research has been conducted in survey articles and tutorials, as evidenced by publications such as [23], to investigate the connections and differences that exist between data warehouses and data lakes. The storage technique of data lakes differs from that of hierarchical data warehouses, as mentioned in [32]. In contrast to the traditional hierarchical data warehouse, which uses folders and files to store information, data lakes have a flatter structure. A data lake uses a system of enhanced metadata tags and a unique identifier for each data element. Data lakes are an alternative to hierarchical data warehouses, which are characterized by strict schema and data manipulation. Data lakes priorities preserving the order in which data arrives and may accept various data forms and sizes without requiring predetermined schema. Data lakes are conceptually large pools of data that collect data in real-time and in the past from many sources, including devices and sensors. This data might be structured, unstructured, semi-structured, or binary in nature. An important idea called schema-on-read states that the data and schema needs are not specified until the data is queried. Data warehouses can be updated, but the data kept there might seem static compared to data lakes, which promote more dynamic data. Several studies [12,27] investigate the ways in which data lakes and data warehouses interact with one another. Although views on their long-term compatibility differ, data lakes and data warehouses are likely to remain side by side for the time being. Data lakes and data warehouses are complementary, which means that data from one can be translated to the other or exchanged between the two. Numerous research opportunities have arisen as a result of this coexistence to investigate the relationship between these two data storage models.

8. Otherno table developments related to data lakes

This work summarizes observations on various interesting directions in data lake applications, despite the fact that it was not primarily intended for applications on data lakes. The goal of delving into these uses is to show how data lakes work and to shed light on some of the key characteristics of data lakes:

■ ***Blockchain technology*****:** The escalating popularity of blockchain technology, as a field focused on decentralized ledgers, is noteworthy. Researchers have delved into the relationship between data lakes and blockchains, as evidenced in studies like [37,38,43]. One specific application of this exploration involves addressing security-related issues within the context of data lakes and blockchains.

■ ***Life science*****:** Instances of data lakes' applications in the life sciences domain are highlighted in sources such as [8,11].

■ ***Smart cities and smart computing*****:** The utilization of data lakes in the context of smart cities and smart computing has been explored and discussed in works like [26,35,36].

9. Our expectation: Towards A Frame work for Data Lake Life cycle Evaluation

In conclusion, the paper's emphasis moves to a thorough evaluation of experts' views on the difficulties and possible future research directions in the data lake area. Author in [20] delves deeply into the difficulties and unanswered questions surrounding data lakes, highlighting the lack of (a) a cohesive idea of data lake architecture, (b) strong data lake governance, and (c) a comprehensive plan for design and implementation. Building upon these findings [45] explores four obstacles that research and implementation of data lakes face. These include keeping up with changing data source structures on integration layers, optimising data processing workflows, cataloguing available data sets, and managing metadata. One of the main goals is to ensure high data quality, particularly in terms of duplicate elimination, in data warehouses and data lakes. Moreover, five significant knowledge gaps are highlighted in [13]: 1) unclear data modelling techniques, 2) no data lake reference architecture, 3) inadequate metadata management, 4) inadequate data lake governance, and 5) no comprehensive strategy for implementation and integration.

The above-mentioned criticisms and insights are much appreciated, since they give useful direction for the future of data lake study and development. While we do our best to have an open mind about the potential for disagreement, we do offer our personal perspectives on current trends in data lake research in this article. The following three points are at the top of our wish list in terms of importance:

■A universally accepted and authorized definition of a data lake is anticipated to emerge in the near future;

■A more comprehensive understanding of the relationship between data warehouses and data lakes is expected to be widely disseminated within the IT community;

■In particular, we are trying to find a way to measure data lakes during their whole lifespan. Data ingestion, maintenance, discovery of related datasets, data interaction, metadata enrichment, data quality improvement, schema evolution, and query-driven data discovery and querying of heterogeneous data are all important parts of this framework that could be modelled after the architecture suggested in [24]. We believe that data lakes are an important topic that might greatly benefit from additional systematic, comparative, and experimental research.

References

[1] Alserafi A, Abelló A, Romero O, Calders T, Towards Information Profiling: Data Lake Content Metadata Management, Proc. IEEE ICDM Workshops 2016.

[2] Alexiou G, Papastefanatos G, Query Driven Entity Resolution in Data Lakes, ISIP 9 May 2019

[3] Amazon, What is governance, risk, and compliance (GRC)? https://aws.amazon.com/what-is/grc/

[4] Bogatu A, Fernandes AAA, Paton NW, Konstantinou N, Dataset Discovery in Data Lakes, ICDE 2020, 709-720.

[5] Che H, Duan Y, On the Logical Design of a Prototypical Data Lake System for Biological Resources, Front BioengBiotechnol. 2020; 8: 553904.Published online 2020 Sep 29. doi: 10.3389/fbioe.2020.553904

[6] Chen Z, Intelligent Data Warehousing: From Data Preparation to Data Mining, 2001.

[7] Cheng Z, Wang H, Li H, Extracting knowledge patterns in a data lake for management effectiveness, EBLDM 2020 E3S Web of Conferences 214, 03045(2020)), https://doi.org/10.1051/e3sconf/202021403045

[8] Couto JC, Borges OT, Ruiz DD, Automatized bioinformatics data integration in a Hadoop-based data lake, CSCP 2022, pp. 137-153.

[9] Dabbèchi, H., Haddar, N.Z., Elghazel, H., Haddar, K. (2021). Social Media Data Integration: From Data Lake to NoSQL Data Warehouse, ISDA 2020

[10] Diamantini C, Potena D, Storti E, A Semantic Data Lake Model for Analytic Query-Driven Discovery, iWAS2021, November 29-December 1, 2021, Linz,Austria, 183-186

[11] Diamantini C, Giudice PL, Potena D, Storti E, Ursino D,An Approach to Extracting Topic-guided Views from the Sources of a Data Lake, Information Systems Frontiers (2021) 23:243–262

[12] Dixon J , Pentaho, Hadoop, and Data Lake (14 October 2010). James Dixon’s Blog. Retrieved Aug. 14, 2022.

[13] Eichler R, Giebler C, Gröger C, Schwarz H, Mitschang B, Modeling Metadata in Data Lakes - A Generic Model, In: Data & Knowledge Engineering (2021), 101931

[14] Farrugia A, Claxton R, Thompson S, Towards Social Network Analytics for Understanding and Managing Enterprise Data Lakes, ACM/IEEE ASONAM 2016

[15] Giebler C, Gröger C, Hoos E, Eichler R, Schwarz H, Mitschang B, The Data Lake Architecture Framework: A Foundation for Building a Comprehensive Data Lake Architecture, BTW 2021

[16] Hai R, Quix C, Jarke M, Data lake concept and systems: a survey. CoRR abs/2106.09592 (2021)

[17] Hammer J, Garcia-Molina H, Widom J, Labio W, Zhuge Y, The Stanford Data Warehousing Project, http://ilpubs.stanford.edu:8090/76/1/1995-10.pdf

[18] Han J, Kamber M, Pei J, Data Mining: Concepts and Techniques (3rd ed.), 2010.

[19] Jemmali R, Abdelhedi F, Zurfluh G, Transferring Relational and NoSQL Databases from a Data Lake, SN Computer Science 3(5) July 2022 DOI: 10.1007/s42979-022-01287-7

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

[21] Langenecker S, Sturm C, Schalles C, Binnig C, Towards Learned Metadata Extraction for Data Lakes, in K.-U. Sattler et al. (Hrsg.): Datenbanksysteme für Business, Technologie und Web (BTW 2021), Lecture Notes in Informatics (LNI), doi:10.18420/btw2021-17

[22] Liu Z, Zhang A, A Survey on Sampling and Profiling over Big Data (Technical Report), Cornell University, 2020.

[23] Nargesianm F, Zhu E, Miller R, Pu K, Arocena P, Data lake management: Challenges and opportunities, VLDB, 2019.

[24] Panwar A, Bhatnagar V, A cognitive approach for blockchain-based cryptographic curve hash signature (BC-CCHS) technique to secure healthcare data in Data Lake, Soft Computing, Nov. 2021. https://doi.org/10.1007/s00500-021-06513-7

[25] Ravat F, Zhao Y, Data Lakes: Trends and Perspectives. International Conference on Database

and Expert Systems Applications (DEXA 2019), Aug 2019,Linz, Austria. pp.304-313. ffhal-02397457

[26] Sawadogo P and Darmont J, On data lake architectures and metadata management, J Int Info Sys, vol 56, pp 97-120, 2021.

[27] Silberschatz A, Korth H, Sudarshan S, Database System Concepts (7th ed.), 2019

[28] Wang L, Exploring Blockchain and Big Data with Alibaba Cloud Data Lake Analytics, Alibaba Clouder August 8, 2018