REVIEW OF RESEARCH JOURNAL

Business Intelligence (AMI23B)

Critical Review Paper: Leveraging the Data Lake: Current State and Challenges - Corinna Giebler , Christoph Gröger , Eva Hoos , Holger Schwarz1 , and Bernhard Mitschang

Reviewed By:

Maha Vajeeshwaran

h21mahna@du.se

Summary:

As essential to the concept of data management, data lake system was built to facilitate for comprehensive and flexible data analytics on large and complex data sets, as a result of the digital transformation, an enormous amount of data is produced, making conventional data warehouse solutions extremely hard to manage. Evidently, various types of data may be evaluated by using data lake, but research is yet to uncover any predefined use cases as a standard requirement. Despite the numerous advantages of the data lake, there are various challenges to its deployment and usage. This 2019 article investigates the present state of data lake concept and explores existing design and realization forms of data lake such as governance and data models. The authors highlighted research gaps and obstacles in data lake architecture, data lake governance, and a comprehensive strategy to implementing the data lakes in operation judging by past studies. The authors observe there were no widely accepted concepts for data lakes, and there are many contradictory definitions. They noticed that a single-source data lake was not very well adopted in data lake. They found that data lake without proper metadata and data governance leads to risk.

The authors ignore the distinction between a data lake and data lake governance, viewing it as an integral component of a data lake. Zone, pond, and lambda architecture are forms of data lake architecture. They identified that in zone architecture data, which assigns to a zone based on the processing degree. The key advantage is that data ta can be viewed as raw data in a raw zone, regardless whether it is available in raw or preprocessed format. In contrast, although if the data is easy to examine and process in pond architecture, once the data leaves the raw data point, it is conditioned, and the original format is lost. The Lambda architecture allows data to separate batch and real-time processing; yet, it is often used in practice. They also noticed that the literature only covered a small proportion of the data lake. The data droplets concept and the data vault were discussed in the context of data modelling. Every document in a data lake is modelled as an RDF graph in the data droplets model. Data vault, but at the other hand, is designed for structured data despite being flexible and simple to model. They observed that many other techniques are only applicable for structured data and provided with no guidelines.

According to writers, metadata management is a critical element of the data lake that records information on actual data, such as schema information, semantics, or lineage, based on existing literature data catalogues utilized to store metadata. They discovered that, despite its importance for data lakes, no clear approach to cover all data lake information and emphasizing the need for more study in this area. Metadata management is merely one aspect of overarching data lake governance, according to this article. As diverse types of data are maintained in the data lake, governance must find the right balance between control and flexibility, and it needs to consider the differences. They observed that neither of the traditional governance models consider the various kinds of data managed in data lake and their governance needs. They argue that data lake governance must satisfy new guidelines such as balancing flexibility and control, as well as the necessity for unique data lake concepts which would be just as important for the data lake management.

Objectives:

The primary goal of this article is to examine the present status of the data lake concept, describe existing design and realization aspects, identify important challenges, and to investigate gaps in leveraging data lake, however, it also extends to a comprehensive literature review which predefine the authors findings.

Contributions:

The authors looked into a number of past studies and addressed the present state of data lake concepts. The authors discovered that data lakes are now being redefined to include data from various sources, and Dixon's central point of strong raw data is reflected in all the definitions examined. They proposed that effective management of data lake and storage systems are correctly employed, either on their premises or in the cloud.

The authors claim that various portions of the literature are underrepresented and that there is a lack of a comprehensive strategy for integrating Data Lake. They also observed that heterogeneity of concept causes severe challenges in data Lake architecture, since there is no commonly accepted architecture and some proposed architecture that does not correspond with data lake concepts. As a result, they proposed looking into and analysing the current alternatives in order to find similarities and flaws. They argue that data lake architecture should be generalized and comprehensive.

The findings reveal that there is no guidance on how to utilize alternative models for data lakes, and that many of the existing models are solely available for structured data. It highlighted the need for more study in order to build models that can close these gaps. Due to the lack of a metadata management strategy that addresses all data lakes, they argue that the need for more investigation into alternatives apart from data lakes cannot be overlooked.

The author discovered that the data lake lacked a comprehensive integration strategy, and that this strategy takes into account the interdependencies between various data lake aspects such as data lake architecture and data lake modelling, which combine into a single comprehensive and systematic data lake concept. They also suggest that further study should be performed in these areas such that data lakes can be used in practice more efficiently.

Strengths:

By exploring more into the prior literature, the authors clearly address the current state and major challenges in leveraging the data lake. They found the missing proper strategy to realize data lake by analysing the current literature. They also examined the similarities and shortcomings in existing data lake architecture, as well as the need for a more general and comprehensive data lake architecture. Furthermore, they tried to fill gaps and flaws in the existing literature, stating that further study on data lake governance is necessary to make it more flexible and accessible to the public. And it is suggested that more study be done on data lake architectural, such as modelling, metadata management, and data governance in order to develop a holistic data lake concept that can be used in practice.

Weakness:

The weakness of the paper is shown in how it addresses the current situation and key challenges in data lake management. It fails to provide further literature with regard to related studies, eg. RMU University professors released a study paper in 2018 that looks further into the problems and possibilities involved in constructing data lakes (Anne Shepherd et al: 2018, Volume 19). Despite some similarities to the study report, this article focuses on data lake trends and perspectives (Franck Ravat et al: 2019). It is necessary to provide further analyses and suggestions in order to do future study on this subject. Finally, this article is outdated, and sophisticated data lake solutions such as AWS data lake, Cloud data platform, Google data lake, Azure data lake, Snowflake data lake, and many others have dominated the market.

Conclusion:

Overall, this article examines the present state of the data lake and underlines the obstacles that many businesses face by using it, specifically, in data management, as such shows that data lake system was built to facilitate for comprehensive and flexible data analytics on large and complex data sets which even helps enterprise operation optimization.

It also looks at how to create a data lake approach by looking at a variety of previous literatures. However, I believe that this study provided insights into how to enhance data lake design, data lake modelling, metadata management, and data lake governance, but it did not specifically address how to do so. Furthermore, this research focuses on relatively old literature rather than sophisticated data lake solutions that are currently on the market. To deploy enhanced data Lake solutions, more study is needed to gain clear insights regarding existing utilized data lake solutions.

Reference:

1. Anne Shepherd, Chalermpon Kesa, James Cooper, Joel Onema, Paul Kovacs: Opportunities And Challenges Associated With Implementing Data Lakes For Enterprise Decision-Making, Issues in Information Systems, Volume 19, Issue 1, pp. 48-57, (2018)
2. Franck Ravat, Yan Zhao: Data Lakes: Trends and Perspectives,HAL Id: hal-02397457, (2019), https://hal.archives-ouvertes.fr/hal-02397457