Literature Review on Hybrid Data Architecture (Data Warehouse/ Lake)

Seminar paper

Ollmann, Henry, FH Wedel, Wedel, Germany, imca104224@fh-wedel.de

Abstract

The past decades organizations experiencing exponential growth of data. After all, data is the key to being among the winners in the digital transformation and optimising business process along the entire value chain. Common approaches to handle this data were a data warehouse and data lake. However, these approaches face too many challenges with today's data. In order to eliminate those challenges a new promising hybrid data architecture is introduced: a data lakehouse.

The main objective of this systematic literature review is to identify the current literature on Data Lakehouses. Besides that, this paper defines success factors for implementing a data lakehouse. This paper shows the relevance of hybrid data architectures and highlights the need of research to specify the implementation and best practices of data lakehouses.

Keywords: Data Lakehouse, Hybrid Data Architecture, Data Management, Systematic Literature Review.

Table of Content

Introduction

Information has increasingly become one of the most important resources a company can have. This includes customer information, process information, product information, competitor information, and any other information related to the field in which the company operates. As our society has made a shift toward digital media, more data sources have become available that serve as “oil” (Shiyal, 2021) to companies. As years passed, the rate at which data is captured has accelerated, leading to the importance of data architecture to efficiently use the data. Overall, the great challenge is that managing data gets necessarily complex, because of the ever-growing complexity data (Begoli et al., 2021). Armbrust et al. (2021) point out four key problems with the current data architecture:

- Reliability
- Data Staleness
- Limited support for advanced analytics
- Total Cost of ownership

To overcome these problems, a hybrid data architecture is established called Data Lakehouse, which basically combines a Data Warehouse and Data Lake. So far, there are only a few different approaches available. In the past several decades, data science and analytics have played a significant role in strategic growth for many organizations (Miladinović et al., 2022). That is why there is significant and increasing interest in a Data Lakehouse architecture, however a lack of systematic analysis and review on the current literature is apparent on this topic.

This research constitutes a relatively new area which has emerged from Data Warehouses and Data Lakes. The objective of this paper is to collect and combine relevant literature of Data Lakehouses and Hybrid Data Architecture. Besides that, I hope that my findings will be useful for organizations that are considering implementing a data lakehouse architecture, as well as for researchers who are interested in further studying in this topic. Considering these objectives, the following questions will guide this part of the research:

What is the current state of the literature?

What are success factors for implementing a data lakehouse?

To answer these research questions, this paper is organized as follows: In the first section of the paper, I provide a basic knowledge on data management, including data types and ACID properties. This section is designed to provide context and background information for the rest of the paper. In the second section, I describe the methodological approach that I have used for my review of the literature on hybrid data a rchitecture and data lakehouse. This includes a description of the criteria I used to select the literature for my review, as well as an overview of the key findings from the selected literature. The third section of the paper focuses on the understanding, implementation, and benefits of a data lakehouse. My conclusion is drawn in the last section.

1 Background

Hybrid data architecture combines two different data management designs, which lead to a Data Lakehouse. As I mention some characteristics later in this paper, I want to give you some basic knowledge concerning these aspects. The following section describes different data types and ACID properties.

Data Types

Data can be classified in three different data types: structured data, semi-structured data, and unstructured data. Depending on the data type its beneficial for a specific data architecture.

Structured data is highly organized and stored in predefined format (Shiyal, 2021). For humans and machines, structured data is comparatively easy to interpret. Usually, the data is stored in database systems that follow tabular format.

A much bigger percentage of all the data in our world is unstructured data. It is not organized and does not conform to a specific format (Shiyal, 2021). Examples for unstructured data are documents, image files, audio files or video files. The lack of structure made unstructured data more difficult to search, manage and analyse. However, the importance of unstructured data is increasing during the past years because it’s a key success factor for data-driven decision making.

Beyond structured and unstructured data, there is semi-structured data, which is a mix of both. Typically, the data is stored in XML files or JSON documents. The difference to structured data is, that it uses tags to define different data points instead of names or columns (Shiyal, 2021).

ACID Properties

ACID stands for Atomicity, Consistency, Isolation, and Durability. These are four key properties of a database transaction. Shiyal (2021) describes these acronyms as following:

- Atomicity → The entire transaction takes place at once or doesn’t happen at all
- Consistency → The database must be consistent before and after the transaction
- Isolation → Multiple Transactions occur independently without interference
- Durability → The ability of a transaction to survive permanent failures

Together, these properties ensure that database transactions are reliable, predictable, and able to maintain the integrity of the data within a database.

2 Literature Review

A literature review is a survey of scholarly sources on a specific topic. It provides an overview of the current state of knowledge on the topic and identifies any gaps or inconsistencies in the existing research. This paper follows the methodology of a systematic literature review based on Fettke (2006), Webster and Watson (2006). In this case, the topic is “Hybrid Data Architecture”, also known as “Data Lakehouse Architecture”, which es a new approach to data management that combines the benefits of data lakes and data warehouses (Armbrust et al., 2021). The following section explains how the literature was filtered and selected. In addition to that, I am going to show a concise overview of the selected literature.

2.1 Selection Process

The methodical approach for this literature review is shown in Figure 1. In total there are five main steps to guarantee the highest quality for a literature review.

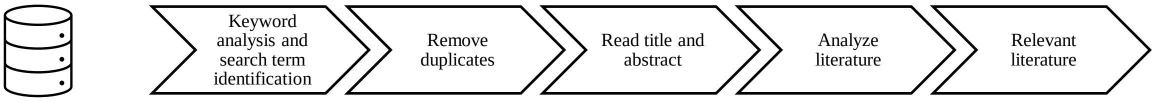

Figure 1 – Selection Process

First, heterogenous databases needs to be identified. To have a high scientific standard on literature the databases ScienceDirect, Beluga and IEEE Xplore were selected. A closer look to the findings on “Hybrid Data Architecture”, however, reveals several gaps and shortcomings. That is why Google Scholar was added to the databases to ensure a broad overview on existing literature. In order to identify relevant literature, an optimal search term needs to be identified. Therefore, I started off with my keyword analysis by searching for “Hybrid Data Architecture” in Google Scholar. A large number of existing studies in the broader literature have applied “Data Lakehouse” as equivalent technical term. For that reason, I extended my final search query from *(“hybrid data architecture”)* to *(“lakehouse architecture” OR “data lakehouse architecture” OR “data lakehouse”)*. Besides Lakehouse Architecture I also wanted to cover some basic understanding on previous architectures. For this reason, I also used the search string *(“data warehouse”) AND (“data lake”) AND (“data lakehouse”)*. That assures me that I can point out the similarities and differences of these data architectures. In the following table, I present my literature findings in the different databases with each search string that was used.

Search String	ScienceDirect	Beluga	IEEE Xplore	Google Scholar
("hybrid data architecture")	3	0	5	31
("lakehouse architecture" OR "data lakehouse architecture" OR “data lakehouse”)	7	4	10	125
("data warehouse") AND ("data lake") AND ("data lakehouse")	5	2	4	63

Table 1 - Search Strings and Number of Results

Before the analysis started, all (34) duplicates had to be deleted. From this initial set of 225 documents, a first analysis was obtained by reading the title and abstract from each work. The main objective here was to identify out-of-scope works, which includes research that does not mention my defined search strings in the article and abstract. Unfortunately, there haven't been many papers that included my search term in their article or abstract. That's why I took some more literature into account that mentioned my search term at least in the introduction. Due to these requirements, a substantial set of 196 works were rejected in this phase. The remaining 39 works were fully analysed to confirm that the documented research adds value to the scope and objectives. By reviewing the citations and references I was unfortunately not able to identify more literature. However, I was able approve that the literature I found was relevant because the literature I found was referenced by other authors. In total 9 relevant papers were identified.

2.2 Selected Literature

After the selection process was done, the condensed number of studies are presented in Table 2. This literature review concluded with 9 papers. The selected literature is categorized in four distinct categories: "Features and characteristics", "similarities and differences", "implementation" and "Benefits & Challenges". According to these categories I analysed the selected literature.

Author	Defintion & Characteristics	Similarities & Differences	Implementation	Benefits & Challenges
Armbrust et al., 2021	●	●	●	●
Azeroual et al., 2022				●
Begoli et al., 2021	●	●	●	
Bureva, 2020	●	●	●	
Miladinović et al., 2022	●	●	●	
Nambiar & Mundra, 2022	●			
Oreščanin & Hlupić, 2021	●	●	●	
Shiyal, 2021	●	●		●
Tovarňák et al., 2021	●		●	

Table 2 - Literature Findings

All literature that has been found on this topic has been published in the last two years. This shows that hybrid data architecture, especially data lakehouses, are still new concepts that need more development and research. On the other hand, it shows the growing relevance of this topic.

3 Results

This section summarises the findings and contributions made in four subsections. In the following, I will first define a Data Lakehouse and show the characteristics of this architecture. Then, I will briefly emphasise on the differences and similarities between the established data architectures (data warehouse + data lake) to the Data Lakehouse. In the next chapter I point out benefits and challenges of this new data management concept. The third section brings the knowledge of implementing a data lakehouse together and gives success factors for an implementation. Finally, I point out benefits and challenges of this new concept.

3.1 Definition of a Data Lakehouse

A Data Lakehouse is a hybrid approach that takes the best concepts from both, the Data Warehouse and Data Lake, and puts them together while trying to eliminate the downsides of both models (Bureva., 2020).

A data warehouse is a repository of integrated data that is used for reporting and analysis (Nambiar & Mundra, 2022). It typically uses pre-defined schema to ensure that the data is clean and consistent, making it easier to query and analyse the data using tools like SQL. A data lake, on the other hand, is a storage system that hold raw data in its native format until it is needed (Bureva, 2020). This makes it ideal for storing large amount of unstructured data, such as logs, sensor data, and social media posts. In the last decade years Data Lakes have been considered to be able to replace Data Warehouses (Oreščanin & Hlupić, 2021) and became increasingly popular as a solution to analyse and store heterogenous data (Begoli et al., 2021).

A data lakehouse combines the advantages of both systems, allowing users to store and analyse both structured, semi-structured, and unstructured data in a single platform (Oreščanin & Hlupić, 2021). Data lakehouse are equipped with support for schemas, capabilities for reading and writing data (Bureva, 2020).

A popular explanation among the literature of a data lakehouse is that it is a low-cost and directly accessible storage that also provides traditional analytical DBMS and performance features like ACID transactions, data versioning auditing, indexing, caching, and query optimization (Armbrust et al., 2021). Shiyal (2021) identified several characteristics of a data lakehouse. They include support for various data types, such as structured, semi-structured and unstructured data; the ability to support important AI workloads like data science, machine learning, deep learning, and natural language processing; the ability to scale storage and compute resources independently; the ability to use data efficiently by storing it in multiple formats; and the ability to offer direct access to the source data for data analytics and BI tools. In addition, Shiyal (2021) agrees to Armbrust definition, that a data lakehouse also supports ACID transactions.

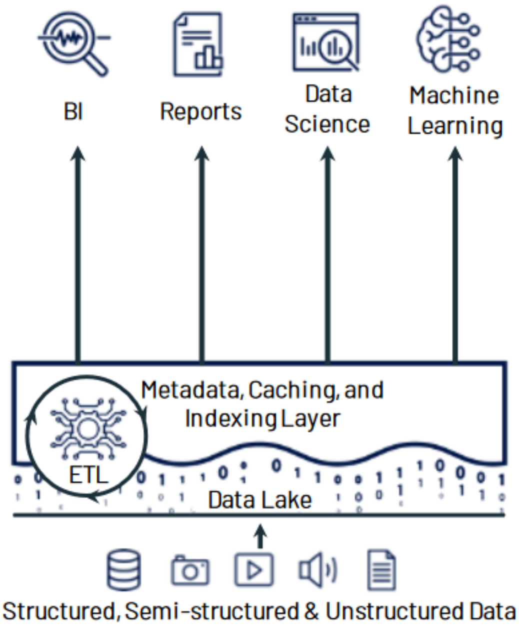
BI
Reports
Data Science
Machine Learning
Metadata, Caching, and Indexing Layer
ETL
Data Lake
Structured, Semi-structured & Unstructured Data

Figure 2 - Data Lakehouse by Armbrust et al. (2021)

Batch and streaming ETL/ELT jobs are used to continuously normalize, transform, and curate the data. Structured access to the data can be achieved through distributed SQL query engines, which allow uses to query the data using SQL and other familiar tools and technologies (Tovarňák et al., 2021).

3.2 Differences between Data Architectures

There are significant differences between the existing data architectures to a data lakehouse. Principally, data warehouses, data lakes, and data lakehouses are all different types of data management platforms that are designed to support the storage and analysis of large amounts of data. Each platform has its own unique capabilities and is intended for different uses, depending on the type and amount of data being managed and the specific needs of the organization. Table 3 provides a comparison between Data Warehouses, Data Lakes, and Data Lakehouses using eleven parameters which is based on the combined knowledge of Shiyal (2021), Miladinović et al. (2022), Nambiar & Mundra (2022) and Oreščanin & Hlupić (2021).

Parameters	**Data Warehouse**	**Data Lake**	**Data Lakehouse**
Data	Business Processes	Raw data	Historical data as well as real-time data
Data Type	Structured Data	Semi-structured and unstructured data	Structured, semi-structured and unstructured data
Data Granularity	aggregated level of detail	Data at low level of detail	Data at all leves of detail
Schmea	Schema-on-read	Schema-on-write	Schema-on-read Schema-on-write
Data Processing	Time-consuming to introduce new content	Helps with fast integration of new data	Takes advantage of distributed computing architecture
Storage	Expensive for large data volumes	Designed for low-cost storage	Designed for low-cost storage
Agility	Less agile, fixed configuration	Highly agile adjustable configuration	Highly agile, adjustable configuration
Security	Mature	Maturing	Both mature and maturing
Purpose	Data analytics and business intelligence	Machine learning and artificial intelligence workloads	Analytics and machine learning workloads
Tools	Mostly various commercial reporting, BI, data analytics, and data visualization tools	Can use open-source tools such as Hadoop or Map reduce	Cloud-based products that combine all components into one single application
Users	Business Professionals	Data Scientists et. al.	Whole business environment

Table 3- Differences and similarities of data architectures

As shown, Data Lakehouses abilities are similar to the existing architectures, but it offers several new features that set it apart from others (Oreščanin & Hlupić, 2021).

3.3 Success factors for implementing a Data Lakehouse

The following section summarises important factors to implement a data lakehouse successfully. By merging the knowledge of the different literatures together, following success factors were elaborated:

1. Identify the goals and objectives of the data lakehouse and develop a clear plan for achieving them. This may involve conducting a data audit to identify the types of data that will be stored in the data lakehouse and defining the uses cases and business value that the data lakehouse is expected to support (Miladinović et al., 2022).

2. Choose the appropriate technologies and tools for the data lakehouse, considering factors such as scalability, flexibility, efficiency, and integration with other systems. This may include selecting a data lakehouse platform, as well as tools for data ingestions, transformation, and access (Oreščanin & Hlupić, 2021).

3. Design the data architecture for the data lakehouse, including the logical and physical data models, data flows, and data pipelines. This should be based on the identified goals and objectives and should support the chosen technologies and tools (Armbrust et al., 2021).

4. Develop and implement processes for data governance and management, including data quality, security, and access control (Begoli et al., 2021). This should include establishing roles and responsibilities, as well as defining policies and procedures for data ingestion, transformation, and access.

5. Populate the data lakehouse with data, and test and validate the data lakehouse to ensure that it is working as expected. This may involve importing data from existing sources, as well as ingesting new data in real-time (Bureva, 2021).

6. Provide training and support to users of the data lakehouse and monitor and evaluate its performance over time. This may include regularly reviewing the data and its usage, as well as adjusting and improvements as needed (Begoli et al., 2021).

Implementing a Data Lakehouse requires careful planning and a deep understanding of the business requirements and goals. It may also require specialized expertise in data management, data processing and data governance (Shiyal, 2021). As such, it may be necessary to work with a team of experts or a specialized vendor to successfully implement a Data Lakehouse.

3.4 Benefits and Challenges

A Data Lakehouse offers many benefits as is brings the best of a Data Warehouse and Data Lake together. Despite all its benefits this maturing data management system still faces many challenges. This section highlights the benefits of a Data Lakehouse in contrast with its challenges.

One key advantage of Data Lakehouse is its ability to support real-time analysis of data. Because data is stored in its native format, it can be accessed and processed in real time, allowing for real-time analysis and decision making. This is a significant improvement over traditional data warehouses, which often suffer from data latency and require data to be pre-processed before it can be analyzed.

Another advantage of Data Lakehouse is its ability to support a wide range of data types and formats (Miladinović et al., 2022). Unlike traditional data warehouses, which are limited to structured data, Data Lakehouse can handle a wide range of data types, including unstructured data, semi-structured data, and structured data (Shiyal, 2021). This allows organizations to store and manage a wide range of data, including text, images, audio, and video.

According to Shiyal (2021) and Armbrust et al. (2021) the implementation of data lakehouse will result in cost savings due to streamlined processes and the use of cheap blob storage.

Despite these advantages, there are also some potential challenges and limitations to consider when using a Data Lakehouse. For example, because Data Lakehouse stores data in its native format, it can be more difficult to manage and maintain than a traditional data warehouse, which has a well-defined schema and data model (Miladinović et al., 2022). Additionally, Data Lakehouse may require specialized expertise and tools to extract value from the data, which can be a barrier for some organizations (Shiyal, 2021). According to Azeroual et al. (2022) the key disadvantage of data lakehouse is its “inmaturity”.

Overall, Data Lakehouse offers many benefits over traditional data lakes and data warehouses, including real-time analysis and support for a wide range of data types and formats. However, it also comes with some challenges and limitations, and may not be the right solution for every organization. As such, it is important to carefully consider the business requirements and goals before deciding to implement a Data Lakehouse.

4 Discussion

This paper has practical implications that are important to different scenarios.

In academic research, this literature review can provide the foundation for a new study as I highlight the key findings, theories, and methods of previous studies in the field of hybrid data architecture and data lakehouses. It can also help researchers to identify gaps in the existing knowledge and to develop hypotheses for their own research.

In policy making, I can help to inform decision-making by providing an overview of the current state of knowledge on data lakehouses. It can help policymakers to identify the key benefits and challenges.

In the business world, this paper can help organizations to identify trends and developments in their industry, and to develop strategies for staying competitive and innovative. It also can help as a brief guidance on implementing a data lakehouse.

The field of data architecture is constantly evolving, and there are several key trends that are likely to shape the future of data architecture. These trends include:

1. The growth of cloud-based data architecture: As more organizations move their data and applications to the cloud, there is likely to be a shift towards cloud-based data architecture. This will enable organizations to take advantage of the scalability, flexibility, and cost-efficiency of the cloud, while also enabling them to easily access and analyze data from multiple sources.

2. The growing importance of data governance and data quality: As organizations collect and process more data, there is an increasing need for effective data governance and data quality management. This will involve implementing policies and processes to ensure that data is accurate, consistent, and accessible, and that it is used ethically and responsibly.

3. The emergence of new technologies and tools: New technologies and tools are constantly emerging that are enabling organizations to collect, process, and analyze data more efficiently and effectively. Examples include machine learning, natural language processing, and data visualization tools. These technologies and tools will likely play a key role in the future of data architecture.

Overall, the future of hybrid data architecture is likely to be shaped by the growth of cloud-based solutions, the growing importance of data governance and data quality, and the emergence of new technologies and tools.

I am aware that my research may have four limitations. The first is that the selection of my reviewed literature differ in quality and relevance concerning hybrid data architecture. Some sources I used in the literature review are not as relevant or of high quality as some other papers. The second is that my literature review can only provide a snapshot of the current state of knowledge on hybrid data architecture. The data lakehouse is a concept that is still being developed and refined. It does not yet have a definitive form or ready-made solution and is therefore a work-in-progress that is constantly evolving (Begoli et al., 2021). New research and developments may have emerged since the literature review was conducted, which could have affected my conclusions. Another disadvantage regarding my methodology is that it only considers my perspective which may not provide a comprehensive or balanced view of the current state of knowledge of this topic. Other researchers may have different perspectives on data lakehouses that could provide valuable insights and contribute to a more comprehensive understanding. My last limitation is that I couldn’t consider all papers as some papers were not accessible or did not fullfill the requirements of a scientific paper.

5 Conclusion

The goal of this literature review was to analyse the current state of literature on hybrid data architecture and to provide a guidance on implementing one approach: a data lakehouse. My research has highlighted the ever-growing importance of a hybrid data architecture. To summarize, my work investigated the characteristics of a data lakehouse and gives a brief overview on this architecture. Besides that, I have outlined the differences and similarities of a data warehouse, data lake and data lakehouse. I have devised success factors for implementing a data lakehouse. Finally, I investigated the benefits and challenges of a data lakehouse. All in all, it can be said that the data lakehouse is a promising hybrid data architecture that might eliminate the downsides of data warehouse and lake. However, this is still a new approach and is in an early phase of their development and is yet to mature. In the long run this architecture needs more use cases and best practises for its implementation as well as usage.

On this basis, I conclude one piece to the bigger picture of the research on hybrid data architecture and data lakehouses, which adds value to scientific research and organisation who eventually want to implement a data lakehouse by their own. My work clearly has some limitations. Nevertheless, I believe my work could be the starting point for guiding organisations on implementing a data lakehouse and further research.

References

Watson, R. T., and Webster, J (2002) ‘Analyzing the past to prepare for the future: Writing a literature review’, MIS Quarterly, pp. xiii–xxiii.

Fettke, P. (2006) ‘State-of-the-Art des State-of-the-Art’, Wirtschaftsinformatik, pp. 257–266.

Armbrust, M. *et al.* (2021) ‘Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics’, p. 8.

Azeroual, O. *et al.* (2022) ‘Combining Data Lake and Data Wrangling for Ensuring Data Quality in CRIS’, *Procedia Computer Science*, 211, pp. 3–16. Available at: https://doi.org/10.1016/j.procs.2022.10.171.

Begoli, E., Goethert, I. and Knight, K. (2021) ‘A Lakehouse Architecture for the Management and Analysis of Heterogeneous Data for Biomedical Research and Mega-biobanks’, in *2021 IEEE International Conference on Big Data (Big Data). 2021 IEEE International Conference on Big Data (Big Data)*, pp. 4643–4651. Available at: https://doi.org/10.1109/BigData52589.2021.9671534.

Bureva, V. (no date) ‘INDEX MATRICES AS A TOOL FOR DATA LAKEHOUSE MODELLING’, p. 25.

Miladinović, D., Popović, J. and Korolija, N. (2022) ‘The Evolution of Big Data Analytics Solutions in the Could’, p. 6.

Nambiar, A. and Mundra, D. (2022) ‘An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management’, *Big Data and Cognitive Computing*, 6(4), p. 132. Available at: https://doi.org/10.3390/bdcc6040132.

Oreščanin, D. and Hlupić, T. (2021) ‘Data Lakehouse - a Novel Step in Analytics Architecture’, in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO). 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, pp. 1242–1246. Available at: https://doi.org/10.23919/MIPRO52101.2021.9597091.

Shiyal, B. (2021) *Beginning Azure Synapse Analytics: Transition from Data Warehouse to Data Lakehouse*. Berkeley, CA: Apress. Available at: https://doi.org/10.1007/978-1-4842-7061-5.

Tovarňák, D., Raček, M. and Velan, P. (2021) ‘Cloud Native Data Platform for Network Telemetry and Analytics’, in *2021 17th International Conference on Network and Service Management (CNSM). 2021 17th International Conference on Network and Service Management (CNSM)*, pp. 394–396. Available at: https://doi.org/10.23919/CNSM52442.2021.9615568.