Mednarodno inovativno poslovanje
Journal of Innovative Business and Management

Data Lakehouse:
Benefits in small and medium enterprises

Darko Golec*

Abstract: This article argues that the collection and processing of structured and unstructured data is necessary in small and medium-sized enterprises to advance in the business. The architecture, as it is known today, of a Data Warehouse is superseded by the larger Data Lakehouse. It is based on open data formats with direct access and has excellent support for machine learning and data science. To the enterprises that currently process only structured data, so they are technologically behind the competition, Data Lakehouse can help to solve big challenges based on Data Warehouses and Data Lakes, including data obsolescence, unmanageable data, and misplaced data. In the article it is presented how small and medium-sized enterprises are choosing to make a technological shift in the direction of Data Lakehouse, which helps them with strategic decisions and further steps in the enterprise for faster growth and development. An article may be the basis for a specific case study, making in-depth and detailed examination of the benefits in a particular case.

Keywords: Data Lakehouse, Data Warehouse, Data Lake, Data Analysis, Enterprise

JEL: M21

*dr., IBM Slovenija in Doba Fakulteta, Maribor, darko.golec@gmail.com

©Copyrights are protected by = Avtorske pravice so zaščitene s: Creative Commons Attribution-Noncommercial 4.0 International License (CC BY-NC 4.0) = Priznanje avtorstva-nekomercialno 4.0 mednarodna licenca (CC BY-NC 4.0)

DOI 10.32015/JIBM.2022.14.2.5

Mednarodno inovativno poslovanje = Journal of Innovative Business and Management

ISSN 1855-6175

Data Lakehouse:
Prednosti malih in srednje velikih podjetij

Povzetek: Članek zagovarja dejstvo, da zbiranje in obdelava strukturiranih ter nestrukturiranih podatkov imata velik pomen za napredek oziroma uspešno poslovanje malih in srednje velikih podjetij. Arhitektura Data Warehouse (podatkovno skladišče), kot jo poznamo danes, je nadomeščena z večjo, ki so jo poimenovali Data Lakehouse. Temelji na odprtih podatkovnih formatih z neposrednim dostopom in ima odlično podporo za strojno učenje in podatkovno znanost. Podjetjem, ki trenutno obdelujejo le strukturirane podatke in tehnološko zaostajajo za konkurenco, lahko Data Lakehouse pomaga rešiti velike izzive, ki temeljijo na podatkovnih skladiščih in podatkovnih jezerih, vključno z zastarelimi, neobvladljivimi in izgubljenimi podatki. V članku je predstavljen način, na katerega se mala in srednje velika podjetja odločajo za tehnološki premik v smeri implementiranja arhitekture Data Lakehouse, ki jim pomaga pri strateškem odločanju in nadaljnjih korakih v podjetju za doseganje hitrejše rasti in razvoja.

Ključne besede: Data Lakehouse (združeni tehnologiji podatkovnega jezera in podatkovnega skladišča), podatkovno skladišče, podatkovno jezero, podatkovna analiza, podjetje

1 Introduction

In the article we focus on small and medium-sized enterprises (SMEs) which represent 90% of all enterprises and have more than 50% of all employees worldwide. They produce more than 40% of the gross domestic product (GDP) in the economy (World Bank SME Finance: Development News, Research, Data | World Bank, n.d.). Their employees must be able to integrate increasing amounts of data into their enterprise structure regardless of the source, form, or amount of information. Any data can be good data and bring the key to success for the enterprise or company. While a company is typically an organization engaged in an economic activity for the purpose of earning profits for the stakeholders, an enterprise may not be a formal company in many instances. Data processing plays a key role in modern society and represents an important benefit in the business world. SMEs are more exposed to financial factors as well as more vulnerable in other areas and are not robust enough to withstand the onslaught of economic and global competition from larger enterprises. However, they are an integral part of various operational and strategic processes that support and build a better tomorrow.

To remain competitive, SMEs must stay in touch with technology and use all possible data to figure out the appropriate strategy and be able to rely on different sources of information. Business Intelligence (BI) is a set of methodologies, processes, architectures, and technologies that transform processed and raw data into meaningful and useful information that enables users to make a variety of decisions that are well thought out and based on real data in real time.

Enterprises rely on Data Lakes and Data Warehouses to extract relevant data. A Data Lake stores data in a raw, unprocessed form and is only transformed if formatting is required for further use, while Data Warehouses store processed and structured data that is intended for decision-making processes in the enterprise. The combination of the two systems is a Data Lakehouse commonly used by the term DLH. The idea is about combining all types of data into one whole. A DLH is designed to process large amounts of structured and unstructured data, which makes it possible to achieve future forecasting, machine learning and the correctness of strategic decision-making in companies. This concept is particularly suitable for cloud environments with a separate processing and storage, where different computing applications can run on demand on separate compute records while directly accessing the same data on an on-premises or cloud system. It is becoming increasingly popular even among SMEs, as it is seen as the third generation of data analytics, which fully utilizes concepts of Data Warehouse and Data Lake, thereby increasing the credibility of the enterprise and reducing the possibility of wrong strategic decisions.

As the DLH is a relatively new concept, SMEs are still quite cautious, because the financial investment can be quite large and if it does not justify the investment, then the concern is in the right place. In the price environment, however, there are large deviations, and it is very difficult to calculate the fixed monthly cost of using the DLH system. It is necessary to know what data is present in the enterprise and what goals should be obtained with the data, which is why SMEs are still considering the usefulness of such a system. The importance of data has never been more important than it is today. Any piece of information can be very important information.

An article does not contain a case study and is based on desktop research and data available on the web and other literature. It is divided into six sections, the second of which provides a basic explanation of what DLH is and what functions it provides. The next section explains the data analytics needs of SMEs, discusses data processing issues and the transition to DLH. Fourth section touches on one of the DLH options, namely Azure Synapse Analytics. The last two sections discuss what are the typical benefits and why it would be a good idea for SMEs to think in the direction of building their own DLH and conclusion at the end.

2 Data Lakehouse, key Features and Components

2.1 From Data to a Data Lakehouse

Data is the biggest asset after people for businesses, and it is a new driver of the world economy (Panwar & Bhatnagar, 1 C.E.). Although it seems that SMEs cannot help themselves with most of the data, this is not the case these days. Data management is one of the most serious challenges faced by organizations (Singh & Ahmad, 2019), (Nargesian et al., 2019). New data is generated every day, and this means more business cases of data analytics usage. More and more companies are wondering why their business is not improving year after year, despite very good products, but the answer is in a strategy that allows them to use the data for data analysis, and those companies that cannot do it, have more difficulty penetrating the market than those that have it.

The Data Warehouse, commonly used by the term DWH or DW, is a system used for collecting structured data, and filtered for individual purposes. It is non-volatile, subject-oriented, integrated, time-variant, and non-operational data, gathered from multiple heterogeneous data sources (Saddad et al., 2020). DWH is most used to produce analyses from enterprise resource planning (ERP), customer relationship management (CRM), or other systems. The digital transformation leads to massive amounts of heterogeneous data challenging traditional data warehouse solutions in enterprises (Giebler et al., 2019).

The Data Lake, commonly used by the term DL, is a large, centralized database of raw data. It is defined as a data landing area for raw data from many sources (Panwar & Bhatnagar, 1 C.E.) and unknown purpose. It contains a huge amount of data but requires expertise to acquire. A rapid growth of unstructured data represents a huge percentage of overall data, which is termed as a Big Data (Thomas & Nair, 2020). It generally comes from transactional systems, Internet of Things, and social media (Sawadogo & Darmont, 2021). Data Lake is one of the arguable concepts appeared in the era of big data (Khine & Wang, 2018). Trends and perspectives of Data Lakes are discussed in (Ravat & Zhao, 2019). A Data Warehouse is used by business users, while Data Lake can be understood by data scientists. By putting these two together, DLH can be presented.

The DLH, is a new concept in data architectures that embodies and integrates well established concepts for the systematic management of disparate, large-scale data (Begoli et al., 2021). The DLH combines the best elements of the Data Lake and Data Warehouse, aggregating structured and unstructured data. Its architecture should be able to deal with currently insurmountable challenges that both Data Warehouses and Data Lakes cannot overcome (Orescanin & Hlupic, 2021). It is a new type of architecture for storing structured and unstructured data. Data may be stored in enormous quantities in one place and are immediately suitable for further analysis. At the same time, DLH provides a data structure and provides data management features, such as those in DWH, by running metadata layers at the top of the DL. This allows SMEs to use one system to access all data for a range of projects, including data science, machine learning and business intelligence. Unlike Data Warehouses, the DLH can store and process many different data at a lower price, and unlike Data Lakes, this data can be managed and optimized for structured query language commonly used by the term SQL performance. Data Warehouse architecture will wither in the coming years and be replaced by a new architectural pattern, the DLH, based on open direct-access data formats (Armbrust et al., 2020).

A crucial component of any data management system, including DLH, is data integration. Data integration is the process of combining data from various sources into a singular format that can be evaluated and applied to decision-making. When referring to DLH, data integration entails moving organized, semi-structured, and unstructured data from various sources where it can be quickly viewed, processed, and controlled. Several techniques, such

as bulk processing, real-time streaming, and shift data capture, can be used to integrate data in DLH. Data is extracted from various sources and put through batch processing so that it can be loaded into the DLH in a predetermined format. While analysing data as it is produced by various sources and putting it into the DLH in real-time is the goal of real-time streaming.

Another popular technique of data integration in DLH is change data capture (CDC). CDC involves only loading the changed data into the DLH as opposed to the entire dataset and capturing changes to data sources in real-time. Next important thing is Data quality, which is essential for DLH, and organizations must establish processes and frameworks to ensure that their data assets are accurate, complete, and fit for purpose. Data profiling, data cleansing, data governance, data lineage, and data quality monitoring are all critical components of a data quality. Object storage is usually the foundation for data storing in a DLH. In a storage system known as object storage, data is managed as objects rather than folders in a hierarchical file hierarchy. Data is kept in object storage systems as objects, each of which has a special identification, information, and the data itself. These items can be easily and quickly retrieved because they are kept in a flat address area. Due to the vast amounts of data kept in the system and the variety of data sources, data governance is especially crucial. Defining rules and processes for managing data within the system, including how to classify data, who has access to it, how to handle it, and how to safeguard it from unauthorized access or change, is a necessary step. A strong data governance structure must be established in order to define the norms and rules for data integration, quality, security, and privacy. This will ensure that the data analysis process in a DLH is effective and efficient. The duties and obligations of the various stakeholders, such as data analysts, data engineers, and data scientists, who are engaged in the data analysis process should also be outlined in this framework. In a DLH, effective data representation necessitates a thorough comprehension of the data and its context. The data being displayed must be of a high calibre and must have undergone the necessary processing and cleaning. The intended audience should also be considered when creating the visualization, along with their degree of knowledge and the queries they hope to have answered.

2.2 Data Lakehouse Features

The DLH features are concurrent reading and writing of data, schema support with mechanisms for data governance, direct access to source data, separation of storage and compute resources, standardised storage formats, support for structured, semi-structured and unstructured data, and end-to-end streaming.

Concurrent execution of multiple reading and writing operations, which is especially useful in the production scenario, comes with atomicity, consistency, isolation, durability commonly used by the term ACID, ensuring the same atomicity and consistency of the data specified in the Data Warehouse (Data Warehouse vs. Data Lake vs. DLH: An Overview of Three Cloud Data Storage Patterns | Striim, n.d.).

Schema enforce helps to ensure that the data types are correct and that the necessary columns are present, which prevents bad data from causing damage to the data. If a lot of Data Warehouses and Data Lakes is used, then it is bound to generate excess data - when the same piece of data is stored in two or more separate locations. Not only it is ineffective, but it can also cause data inconsistency when the same data is stored in different versions in more than one table. The DLH can help consolidate data, remove additional copies of the data, and create a single version of the truth for the company. Structured, semi-structured and unstructured data types are supported.

DLH separates compute resources from storage resources to make storage more flexible, standardized, cost-effective and scalable. DLH supports end-to-end real-time streaming and

input from data sources, allowing smart real-time reporting. It is ideal for various workloads, such as data science, SQL analytics and machine learning for business analytics.

2.3 Data Lakehouse Architecture

The DLH has a layered architecture of Ingestion, Storage, Metadata, application programming interface commonly used by the term API, and Data Consumption layers. The ingestion layer is responsible for collecting data from multiple sources and transferring it to a storage layer, which in turn allows several types of data to be stored. The combination of cloud storage and fast, elastic processing provides an inexpensive and scalable solution for building analytical applications (Syed, 2020). A metadata layer is a common catalogue that provides all data for all objects stored in the DLH and provides users with a management function, including ACID transaction, cache, indexing, and data retrieval. The API layer contains a variety of API that allows all end-users to process and access the data needed for further analytics. The data consumption layer ensures that final data can be used for a variety of things, such as BI report creation, and Machine Learning based on Artificial Intelligence.

DLH architecture typically comprises the components of storage, file formats, table formats, query engines and applications. Object storage is currently available in typical cloud service providers, such as Amazon S3, MS Azure Storage, and Google Cloud Storage. They support all types of data storage and facilitate the required performance and security. These systems can be excitable and cheap, which helps to rationalize costs. Typically, column formats that provide important advantages when reading data or sharing data between multiple systems. Common file formats are Apache Parquet, ORC and Apache Arrow (Data Warehouse vs. Data Lake vs. DLH: An Overview of Three Cloud Data Storage Patterns | Striim, n.d.). These files are stored in object storage.

The table format or table shape is a way of organizing and managing all raw data files. Table formats help abstraction of the complexity of the physical structure of data and allow different engines to run simultaneously on the same data. The format of the DLH table allows transactions to be conducted at the data warehouse level together with the ACID guarantees. A few critical functions of the table format are schema development, SQL expression, time travel and data hash. Apache Iceberg, Hudi and Delta Lake are three popular desktop formats. Table formats provide the specifications and API needed to interact with table data.

Query engine ensures and is responsible for data processing. Some query engines allow to connect to BI tools, such as Tableau, making it easier to report directly data stored in the object store. Querying machines such as Dremio Sonar and Apache Spark work seamlessly with formats like Apache Iceberg to enable robust architecture using commonly used languages such as SQL.

The final component of the DLH are applications that communicate with data. This includes BI tools such as Tableau and Power BI, as well as machine learning frameworks such as TensorFlow, PyTorch, which make it easier for data analysts, scientists, and ML engineers to access data directly.

3 Data Analysis in Small and Medium Enterprises

In SMEs, data analysis is in most cases used for business decisions. It is more like structured data inhere. But if all the data is analysed wisely, it can be the key to success. SMEs can benefit greatly from the use of different types of data analysis techniques based on the

business model. Although SMEs are crucial for every economy, they are lagging far behind in the usage of Big Data analytics (Maroufkhani et al., 2020). SMEs are shifting to a data management concept called the DLH, which implements the functionality of structured data warehouses on top of unstructured data lakes (Behm et al., 2022).

Most of the time, statistical data analysis is used in such enterprises (Behm, 2022). It is about researching, collecting, and presenting large amounts of data to detect patterns and trends. Statistical analysis is good if data manipulation is happening for decision-making purposes. To find patterns, trends, and insights, statistical analysis includes gathering, organizing, and analysing data. In order to make wise business choices, it can be used to evaluate structured data, like revenue statistics or customer demographics. Additionally, statistical analysis can be used to assess the success of marketing initiatives or worker productivity. As more companies come to understand the advantages of data-driven decision-making, the use of statistical analysis in SMEs is expanding quickly. Statistical analysis can assist companies in spotting patterns and trends in consumer behaviour that can be used to create more successful marketing efforts. Additionally, it can be used to enhance supply chain operations, which will help companies cut expenses and increase efficiency.

Diagnostic data analysis is a form of advanced analysis that examines data and content to answer the question of why something happened. The characteristic is the downward drilling technique, data detection, data mining and correlation. The data is used to determine the causes of trends and correlations between variables. A diagnostic data analysis is simple and does not require a special specialist in the field of data analytics. With the right tools and techniques, SMEs can perform their own diagnostic analysis, making it a cost-effective solution for small businesses. Finding the underlying cause of a specific problem is not the only goal of diagnostic data analysis. Additionally, it can be applied to avoid future issues. SMEs can take remedial action before an issue gets out of hand by spotting patterns and trends in the data.

Forecast data analysis is used to predict future results based on historical data and statistical modelling, machine learning and data mining. In this case, SMEs must move closer to knowing the basics of Data Analytics and the basics of Artificial Intelligence. Since SMEs frequently have restricted resources and must make important business-affecting strategic choices, this procedure is especially crucial for them.

Prescribed data analysis examines the data and content to provide a recommendation in steps to solve the problem (Salleh, 2018). This format uses a wide range of instruments and techniques, including simulations, graph analysis, complex event processing, recommendation mechanisms and neural networks. In such cases, SMEs need employees who are skilled in data analytics.

SMEs can make a big shift for the better future in general with the help of appropriate data technology. With automation for example. With the help of analysis and data management, enterprises save a lot of available resources and can track what the final proceeds and return on an investment are. For many companies, automation helped with the challenges of the pandemic. Rapid market adjustment, following the guidelines of the introduction of digital technology, not only made it possible to streamline processes by eliminating inefficiencies, but also greatly facilitated the transition to distance work. According to the report (Business Process Automation Market Size, Share and Global Market Forecast to 2026 | MarketsandMarkets, n.d.), the global business process automation market is expected to reach USD 19.6 billion by 2026 with a growth rate of 12.2%. This may mean that even SMEs will experience a revival at the level of automation. SMEs can focus more on the main part of their business and can easily predict where trends are going with all the available data and how much demand is there for a particular thing on the market. In 2020 only 45% of small businesses used data analytics, while only 51% of them thought data analytics was

good. More than 73% of small businesses have prioritized finding new customers, which means they offer open doors in search of new economic commitments.

In the past, performing data analytics involved significant investments in both hardware and software. Additionally, companies needed to have employees - if not a department - dedicated to manually working with data daily, filtering it, and organizing it in the appropriate manner. However, SMEs now have the option of utilizing established cloud systems for their data analysis needs. It is important for these businesses to carefully consider the advantages and disadvantages of using such a system. Generally, cloud platforms offer payment structures based on individual transactions or units of data analysed. If companies have a clear understanding of their objectives and specific data analysis requirements, utilizing a cloud platform can allow them to obtain sophisticated analysis capabilities at a lower cost.

Companies usually begin collecting data from relational databases and third-party APIs online. Data retrieved from databases typically follows a structured format based on predefined tables and fixed ratios. Similarly, data obtained from Web APIs is generally structured or semi-structured, with a defined format structure such as JSON or XML, despite not having a complete understanding of the underlying data structure. In the case of processing unstructured data from sources like video, audio, or other media and linking it to structured data, SMEs may choose to use a DLH instead. The decision of which option to use depends on the specific needs of the business, as there are numerous choices available.

In addition, Data analysis tools should not only offer advantages in terms of functionality, but also be user-friendly and easily accessible. This complexity is addressed through platforms specifically designed for this purpose, which provide a wide range of functionality and the ability to manipulate different types of data. Effective data management is essential to ensure that there are strong policies, processes, roles, and technologies in place to guarantee the availability, usefulness, integrity, and security of data. Additionally, good data management practices should ensure that individuals have access to appropriate, trustworthy data and that it is managed correctly.

4 Data Lakehouse Example with Microsoft Azure Synapse Analytics

In Azure Cloud there are several options for building a DLH – one of them is the use of Azure Synapse (Shiyal, 2021). Microsoft Azure Synapse Analytics is an enterprise analytics service and successor to the Azure SQL data repository. It provides the freedom to query data, using either serverless or dedicated resources at scale (Azure Synapse Analytics - Azure Synapse Analytics | Microsoft Learn, n.d.). With the service, Microsoft aims to expand the modern Data Warehouse and Big Data strategy and enable companies to analyse their data more efficiently and quickly. Azure Synapse Analytics provides tools, which leverage different layers and make transformations. Azure Synapse offers unified analytics platform with tools and capabilities combined to perform the ingest and prepare phases (Figure 1). There is a flexibility to choose the right tool for each job, step, or process without complexity of integrating these tools.

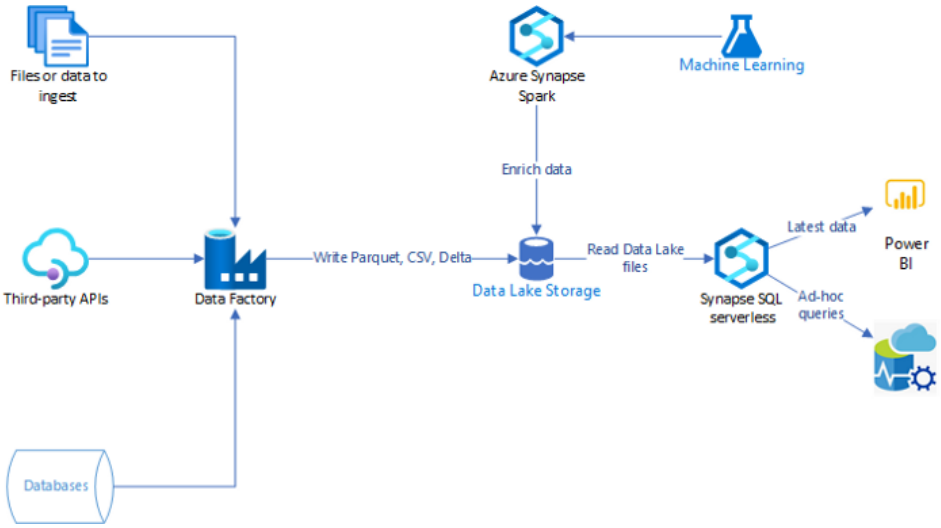

Figure 1. Data Lakehouse Architecture in Azure Synapse.

Resource: Microsoft

To quickly process, prepare, organize, and serve data for urgent business intelligence and machine learning requirements, unified analytics includes a completely integrated workplace for big data and data storage. Azure, which is created to be extremely adaptable, is the foundation for scalability. It is simple to scale up or down to accommodate SMEs' shifting requirements. Strong security features are offered by Azure Synapse Analytics, such as virtual networks and firewalls, Azure Active Directory-based authentication and permission, encryption both in transport and at rest, and network security features.

Performance is based on a distributed design that can rapidly and effectively handle enormous quantities of data. To offer high speed, it makes use of dispersed query processing and data moving. A full DLH solution is provided by integrating Azure Synapse Analytics with other Azure services like Azure Data Lake Storage, Azure Databricks, and Azure Stream Analytics. A variety of tools and services for data preparation, data warehousing, big data analytics, and machine learning are offered by Azure Synapse Analytics, which can assist SMEs in deriving insights from their data and making data-driven choices.

Additionally, Azure Synapse Analytics provides cost-efficient pricing methods, which can be especially helpful for SMEs on a tight budget. The serverless option enables SMEs to only pay for the processing and storage resources they actually use, with no need to maintain or create any infrastructure.

5 Discussion

The benefits of DLH are primarily related to flexibility for storing DL unstructured data while providing the DWH data management tools and features on top of DL, coupling DL and DWH strategically in a larger data management system, and allowing users to leverage the best of both worlds.

The emergence of a solution for collecting, processing, and presenting data in small and medium-sized enterprises is driven by real-world needs. While in the past companies did not place greater emphasis on collecting different types of data, mainly due to a poor understanding of information and different formats, nowadays this data is compatible and available in a wide variety of formats. Solutions for processing all types of data are designed

to process any of the complex data in the range to prepare data for further value-added analyses. Data collection, analysis and visualization are becoming increasingly important and will remain so in the years to come. They enable SMEs to simplify their IT infrastructure and provide a valuable basis for data usage. They also provide valuable support for different types of analytics and provide easier and faster access to large amounts of data.

Data and information are increasing significantly from day to day, so it is important that the data is properly stored, structured, and integrated within the enterprise. The aim of all is to minimize the number of insignificant data or rubbish from the data environment. The concept of DLH contributes to understanding all the different types of information from their location, format, structure, quality, and value. This aspect is crucial for businesses, as it greatly simplifies the entire process of information handling that SMEs are looking for.

The analysis, quality, reliability, and relevance of the individual literature is very important in desktop research. In a few cases, the selected literature is the official website of the service provider (Microsoft Learn, Striim), which confirms the credibility of the data at the highest level. The relevance and quality of the literature can be attributed to the abstracts of international conferences (Roy 2018, Salleh 2019). However, the talk about the consistency of the literature with the written literature is confirmed by most of the other cited literatures.

6 Conclusion

DLH is a relatively new concept that combines the benefits of Data Warehouses and Data Lakes, allowing businesses to store, manage, and analyse vast amounts of data in a flexible and cost-effective way. The primary advantage of DLH is its ability to provide a unified view of data from different sources, enabling businesses to make informed decisions. Another benefit is its high adaptability, enabling the ingestion and handling of both structured and unstructured data. However, it's crucial to assess both the advantages and disadvantages of DLH, including the significant infrastructure and data management tool expenses, which may be expensive for smaller businesses. Furthermore, some companies may find it challenging to adopt DLH because it requires specialized knowledge and skills to manage effectively. Another potential issue with DLH is the potential for data silos to form, limiting access to and capabilities for data analysis. To maximize the benefits of DLH and avoid potential problems, careful planning, investment in infrastructure and tools, and a qualified team of data specialists are necessary. SMEs must also understand the tools correctly, so they can choose the kind that best suits their needs. Nonetheless, DLH offers a promising solution for SMEs to store and process data in a more efficient and cost-effective manner, with the potential to improve market conditions and support future growth in their businesses.

The next step could be a case study based on the application of different types of analysis to SMEs to better clarify whether DLH is appropriate for this sector of the economy. There are some small individual examples of use cases in SMEs but not yet satisfying the general fact about the applicability of DLH.

References

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia Matei. (2020). *Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics – Databricks*. CIDR '21. https://www.databricks.com/research/lakehouse-a-new-generation-of-open-platforms-that-unify-data-warehousing-and-advanced-analytics

Azure Synapse Analytics - Azure Synapse Analytics | Microsoft Learn. (n.d.). Retrieved January 9, 2023, from https://learn.microsoft.com/en-us/azure/synapse-analytics/

Begoli, E., Goethert, I., & Knight, K. (2021). A Lakehouse Architecture for the Management and Analysis of Heterogeneous Data for Biomedical Research and Mega-biobanks. *Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021*, 4643–4651. https://doi.org/10.1109/BIGDATA52589.2021.9671534

Behm, A., Palkar, S., Agarwal, U., Armstrong, T., Cashman, D., Dave, A., Greenstein, T., Hovsepian, S., Johnson, R., Sai Krishnan, A., Leventis, P., Luszczak, A., Menon, P., Mokhtar, M., Pang, G., Paranjpye, S., Rahn, G., Samwel, B., van Bussel, T., … Zaharia, M. (2022). Photon: A Fast Query Engine for Lakehouse Systems. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2326–2339. https://doi.org/10.1145/3514221.3526054

Business Process Automation Market Size, Share and Global Market Forecast to 2026 | MarketsandMarkets. (n.d.). Retrieved January 9, 2023, from https://www.marketsandmarkets.com/Market-Reports/business-process-automation-market-197532385.html

Data Warehouse vs. Data Lake vs. Data Lakehouse: An Overview of Three Cloud Data Storage Patterns | Striim. (n.d.). Retrieved January 9, 2023, from https://www.striim.com/blog/data-warehouse-vs-data-lake-vs-data-lakehouse-an-overview/

Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2019). Leveraging the Data Lake: Current State and Challenges. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11708 LNCS*, 179–188. https://doi.org/10.1007/978-3-030-27520-4_13/COVER

Khine, P. P., & Wang, Z. S. (2018). Data lake: a new ideology in big data era. *ITM Web of Conferences*, *17*, 03025. https://doi.org/10.1051/ITMCONF/20181703025

Maroufkhani, P., Wan Ismail, W. K., & Ghobakhloo, M. (2020). Big data analytics adoption model for small and medium enterprises. *Journal of Science and Technology Policy Management*, *11*(2), 171–201. https://doi.org/10.1108/JSTPM-02-2020-0018/FULL/XML

Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data lake management. *Proceedings of the VLDB Endowment*, *12*(12), 1986–1989. https://doi.org/10.14778/3352063.3352116

Orescanin, D., & Hlupic, T. (2021). Data Lakehouse - A Novel Step in Analytics Architecture. *2021 44th International Convention on Information, Communication and Electronic Technology, MIPRO 2021 - Proceedings*, 1242–1246. https://doi.org/10.23919/MIPRO52101.2021.9597091

Panwar, A., & Bhatnagar, V. (1 C.E.). Data Lake Architecture: A New Repository for Data Engineer. *Https://Services.Igi-Global.Com/Resolvedoi/Resolve.Aspx?Doi=10.4018/IJOCI.2020010104*, *10*(1), 63–75. https://doi.org/10.4018/IJOCI.2020010104

Ravat, F., & Zhao, Y. (2019). Data Lakes: Trends and Perspectives. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11706 LNCS*, 304–313. https://doi.org/10.1007/978-3-030-27615-7_23/COVER

Saddad, E., El-Bastawissy, A., Mokhtar, H. M. O., & Hazman, M. (2020). Lake data warehouse architecture for big data solutions. *International Journal of Advanced Computer Science and Applications*, *11*(8). https://doi.org/10.14569/IJACSA.2020.0110854

Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, *56*(1), 97–120. https://doi.org/10.1007/S10844-020-00608-7/METRICS

Shiyal, B. (2021). Azure Synapse Analytics Use Cases and Reference Architecture. *Beginning Azure Synapse Analytics*, 225–241. https://doi.org/10.1007/978-1-4842-7061-5_10

Singh, A., & Ahmad, S. (2019). Architecture of Data Lake. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2019 IJSRCSEIT |*, *5*(2), 2456–3307. https://doi.org/10.32628/CSEIT1952121

Syed, A. (2020). *The Challenge of Building Effective, Enterprise-scale Data Lakes*. 803–803. https://doi.org/10.1145/3318464.3393816

Thomas, K., & Nair, P. S. (2020). Data Lake: A Centralized Repository. *International Research Journal of Engineering and Technology*. www.irjet.net

World Bank SME Finance: Development news, research, data | World Bank. (n.d.). Retrieved January 9, 2023, from https://www.worldbank.org/en/topic/smefinance

Behm, J., Steinmann, P., & Schlosser, R. (2022). The Data Lakehouse architecture in small and medium-sized enterprises: A use case of diagnostic data analysis. Journal of Business Research, 142, 387-396.

Salleh, R., & Abdullah, Z. (2018). Diagnostic data analytics for small and medium enterprises: A systematic literature review. Journal of Telecommunication, Electronic and Computer Engineering, 10(1-5), 59-63

Roy, S., & Dey, S. (2019). Data Lakehouse: An Architectural Solution for Multi-Structured Data Management. 2019 IEEE 4th International Conference on Computing, Communication and Security (ICCCS), 262-266.