




Requirements on and Selection of Data Storage Technologies for Life Cycle Assessment

Michael Ulbig¹, Simon Merschak¹(✉) , Peter Hehenberger¹ , and Johann Bachler²

¹ Research Centre for Low Carbon Special Powertrain, University of Applied Sciences Upper Austria, Wels Campus, Stelzhamerstraße 23, 4600 Wels, Austria
simon.merschak@fh-wels.at

² AVL List GmbH, Hans-List-Platz 1, 8020 Graz, Austria

Abstract. The importance of a centralized data storage system for life cycle assessment (LCA) will be addressed in this paper. Further, the decision-making process for a suitable data storage system is discussed. LCA requires a lot of relevant data such as resource/material data, production process data and logistics data, originating from many different sources, which must be integrated. Therefore, data collection for LCA is quite difficult. In practice, relevant data for LCA is often not available or is uncertain and has therefore to be estimated or generalized. This implies less accuracy of the calculated carbon footprint. State of the Art research shows that the LCA data collection process can benefit from data engineering approaches. Key of these approaches is a suitable and efficient data storage system like a data warehouse or a data lake. Depending on the LCA use case, a data storage system can also benefit from the combination with other technologies such as big data and cloud computing. As a result, in this paper a criteria catalog is developed and presented. It can be used to evaluate and decide which data storage systems and additional technologies are recommended to store and process data for more efficient and more precise carbon footprint calculation in life cycle assessment.

Keywords: Carbon footprint · Life cycle assessment · Data engineering · Data storage technology

1 Introduction

One of the recent global concerns for politics and economy is the global climate change. This global climate change is caused by the emission of greenhouse gases (GHGs) like carbon dioxide [1]. With 84%, production and usage are responsible for the biggest part of energy related greenhouse gases emissions. The industrial sector makes 90% of the energy consumption of this part [2]. An important indicator for environmental performance of a product is the carbon footprint [1]. For the identification of the carbon footprint of mechatronic systems it is necessary to analyze the whole product life cycle [3]. Ecodesign is defined by the standard ISO/TR 14062 [7] as the integration of environmental aspects into product design and development. The approach is to have the

environmental impacts of a product in mind during its entire life cycle while designing a product. A method for the evaluation of the environmental impacts on all stages of the product's lifecycle is called life cycle assessment (LCA). But the evaluation of the product carbon footprint is rather complex [3]. Especially the limited availability of environmental data across the entire life cycle is significant [4]. Another challenge is, that relevant data is often uncertain in the early design phase of a product [3]. A large number of LCA relevant data sources is already contained in IT-systems of companies like product lifecycle management (PLM) systems and enterprise resource planning (ERP) systems [5]. Other relevant data can be found in 3D CAD-models and technical drawings [6]. This variety of data sources and formats leads to a very time-consuming data acquisition process. It is important to face these problems and to design a system which enables the generation of a CO₂-report based on the available information. Information can be stored in internal systems of a company as well as in external data sources like LCA relevant databases. The objective of this publication is to support the decision-making process for an appropriate data storage technology for the LCA use case. This approach is based on data engineering practices, which enable data-driven decision-making by collecting, transforming, and publishing data. The selection of an appropriate data storage system which should contain LCA relevant data is a first important step for the implementation of an LCA data pipeline. There are many different types of data storage systems but the most popular by literature are the concepts called data warehouse and data lake. The benefit of such an implemented system is, that the data acquisition process would be much faster and cost-efficient. The minimization of the human factor, which frequently results in mistakes, would also promise data of higher quality. In some large and dynamic productions with a high number of products like clothing - or industry 4.0 productions, where manual data inventory is not possible, such a system could be even an enabler. Furthermore, the data centralization would offer more possibilities to estimate missing values via interpolation or machine learning. But this presumes that the data storage system is used and implemented correctly. There is always the danger of malfunctions and technical issues. Additional, since the data is produced by different members of the supply chain, the system is still depending on their contribution. Overall, the system would make LCA more affordable, precise and offer new opportunities for other technologies and LCA approaches. But it must be actively maintained by IT and experts in the field of LCA to ensure overall quality.

2 Background and Related Research

2.1 Life Cycle Assessment

The International Organization for Standardization (ISO) has standardized the LCA method in its basics with ISO 14040 [4] and in detail with ISO 14044 [8]. According to ISO 14040, LCA addresses the environmental aspects and potential impact of the whole product life. This includes the raw material acquisition, production, use and disposal. The general environmental impact assessment requires the consideration of resource use, human health, and ecological consequences. In ISO 14044 LCA is parted into 4 different phases: goal and scope definition, inventory analysis, impact assessment and interpretation.

2.2 Data Engineering

A main task in data engineering is to provide a data infrastructure so that, at the final state, the required data is ready for further analysis like data science [6]. Part of this task is to select an appropriate storage system for the given use case. The data can be stored temporary in staging areas for example. In those, data usually stays only very shortly until it gets used and deleted. Other use cases require to store data in long-term archival storage to be stored for years [7]. Another part of the data infrastructure is a pipeline which extracts data from one system, transforms it and loads it to another system [6]. Recently, two aspects are very common. One is the handling of streaming data like in Internet of Things (IoT) use cases. The other one is handling a large volume of data [7].

2.3 Related Research

There has been research about facing the data acquisition problem of life cycle assessment before. Some projects suggest applying big data as well as cloud computing technologies as solutions. A case study by the German automotive industry came to the conclusion that the use of big data can provide considerable potential to gather and analyze product related data over the entire life cycle [8]. Another study concludes that environmental performance evaluations can profit from big data. One aspect is that big data can support the environmental supply chain. Integrated data resources in the big data era can help to evaluate environmental efficiencies and resource efficiencies in the industrial supply chain [9]. So far, there have been different approaches to the digital support of LCA. Especially in the beef supply chain, efforts have been made to collect and aggregate data over the supply chain via cloud technologies [10, 11].

2.4 Data Pipeline for Life Cycle Assessment

The whole process of data collection, storage and processing is shown in Fig. 1. The required data must be acquired from different LCA relevant data sources via an automated data integration process which centralizes the data. Two very common examples of this integration process are called Extract-Transform-Load (ETL) or Extract-Load-Transform (ELT). The data storage system collects all LCA relevant data automatically and provides them ready to use for LCA methods. The stored data can then be analyzed by different LCA methods. Finally, a carbon footprint report can be generated, and appropriate visualization methods can be used for the display of the assessment results.

3 Overview of Storage Technologies

In the literature, no umbrella term for the terms data warehouse and data lake could be found. Each of the systems mainly stores data and for this reason, the term data storage system is used as group term in this publication.

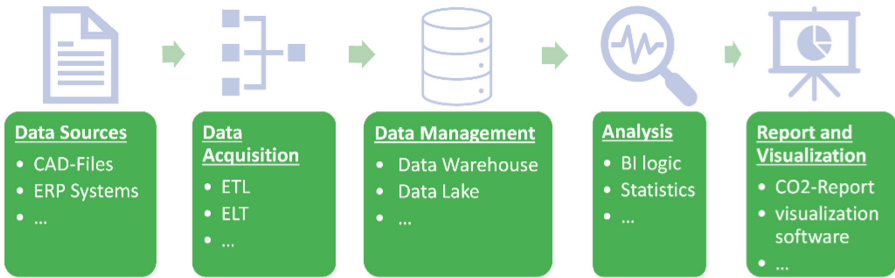


Fig. 1. LCA data pipeline

3.1 Data Warehouse

A data warehouse (DWH) is a logic, centralized data storage system which is separated from operative data storage systems. Ideally, it serves as a company-wide uniform and consistent database supplied by different types of data source systems.

The data flow is shown in Fig. 2. A data warehouse has four characteristics which are subject-orientation, integration, time-variance and non-volatile. Subject-orientation means, that the data management is designed for informing a decision maker about certain subjects. The decision maker should have direct access to the information about the subject. Typical subjects are time, location or product. A schema called Online Analytical Processing (OLAP) orientation can be applied to achieve subject-orientation. The key task of a data warehouse is the integration of data from different operative and external sources. It is important that the collected data must be consistent. Time-variance is the possibility to look at all the data over time. Nowadays also alternative time measures are considered like time transaction. In operative systems the data changes rather often. In contrast to that, the data in a data warehouse should not change after it is successfully integrated to create a history of the values [12].

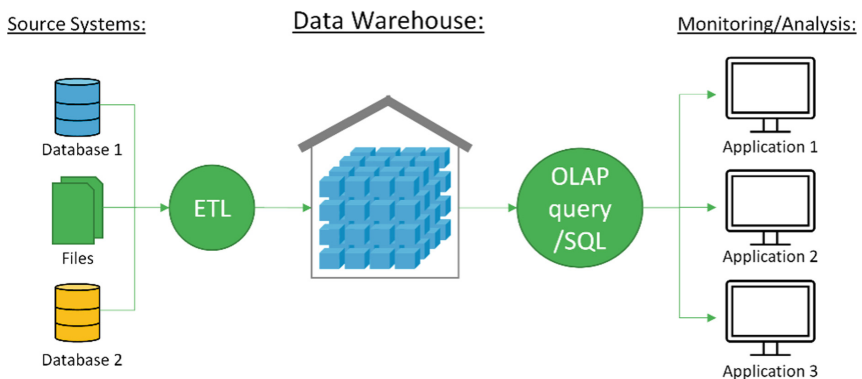


Fig. 2. Data flow for data warehouse

3.2 Data Lake

The main idea of a data lake is to store all relevant enterprise data. In contrast to a data warehouse, the data is stored in the data lake in the raw format. The data flow is shown in Fig. 3. A data lake supposes to include large amounts of data in various data structures. The data lake can be split into two layers. The landing layer contains identical copies of data from the sources systems. This layer is also called mirror layer. The landing layer is the base for the analytical layer. The analytical layer is very dynamic and consists of transformed data from the landing layer. The analytical layer then provides the data ready for business intelligence (BI) applications, analytical models and data visualization [13]. This data can then be accessed via, for example, Spark or structured query language (SQL).

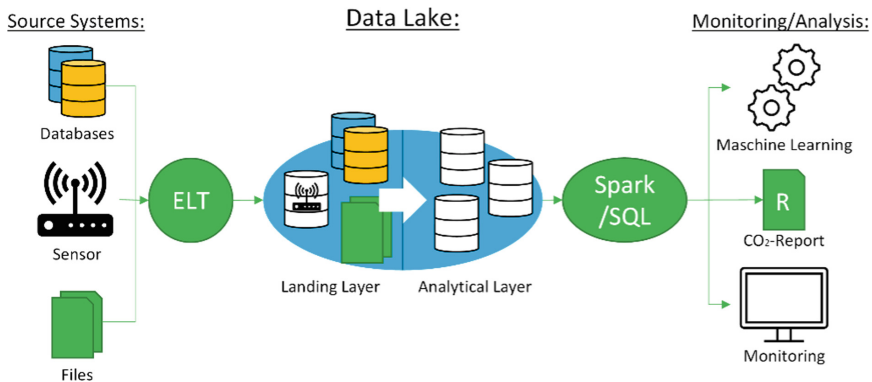


Fig. 3. Data flow for data lake

4 Derived Requirements

There is no standard criteria catalog for the evaluation of different data storage systems. For this reason, different aspects of data storage systems, which are discussed in literature, have been merged to create a criteria catalog. The first aspect which is considered, is the type of data which can be stored in the data storage system. Data can be structured (tables), semi-structured (xml, json) or unstructured (text, pictures, videos) [14]. The second and third aspect are the data in- and output. Looking at the input, there are three different ways how data can be ingested. Those are application, streaming and batch data [7]. For accessing the data storage system there are two methods. One option is to access the data individually and the other is to aggregate information over several datasets. Usual methods for aggregation are sums or medians. Connected to this is the topic of access control and storage duration of data [7]. Another topic which is also very relevant in production facilities, is performance and near-real time analytics. An example for near-real time analytics might be, that a Chief Marketing Officer needs up-to-date sales data of a new product for sales management [15]. The last two topics are big data

and cloud computing. Both have become more and more important, and both have been used in LCA related projects as mentioned in Sect. 2.3. Both aspects are criteria from NIST standards [16, 17]. All relevant criteria are summarized in Table 1. The power consumption of a selected data storage system could also be an additional criterion, but since a solution like this would be integrated in already existing IT infrastructure, it will not be in the focus of this paper.

5 Evaluation and Discussion

5.1 Criteria Applied to Data Warehouse

The technology which is used to implement a data warehouse are relational databases. That means, that a data warehouse is made to store structured data and not semi- or unstructured data [18]. In classical data warehousing the data integration is performed in periodical intervals. The data will then be integrated via an ETL batch process. This process can also include application data from operational systems. Since the data gets integrated via an ETL process, it is not made for ingesting streaming data [12]. The data from a data warehouse can be accessed individually and via aggregation. If a data warehouse is implemented as a schema (OLAP), it is much more made to aggregate data then to access data individually. A data warehouse itself does not provide any access control, but there are extensions which allow such access control [12]. The same applies for the storage duration. A data warehouse by itself can be saved and archived via backups. Besides that, it can also just be stored like a normal relational database. In addition, a data warehouse provides a historicization concept called slowly changing dimension (SCD) [12]. The performance of a classical data warehouse is limited by the speed of the ETL process. So, it would not be possible to achieve near-real time analytics with a classical data warehouse. The challenges which big data is facing, cannot be solved with the use of a relational data warehouse [12]. Since a classical data warehouse would consist of a relational database and could not hold the volume of data which is usually used in big data. This argument can also be applied to the characteristic's velocity and variation. One addition which is rather common is a technique called slowly changing dimension (SCD). SCD has different strategies to historicize semantically or formally changing data which provides variability [12]. Since usually the technology which is used to implement a data warehouse are relational databases, it can be rather simply realized via cloud computing. Data warehouse solutions are hosted by cloud providers like Amazon Web Service (AWS) ® and IBM ® [19, 20].

5.2 Criteria Applied to Data Lake

A data lake stores structured, semi-structured and unstructured data [18]. In concept, the data lake is also able to include operational databases. This means that application data would be includable. The same applies for streaming data and batch data. A data lake has an additional serving layer which provides more efficient and comfortable access [12]. Since it is possible that the types of the stored data are very different as well as the storage technologies which would contain the data, the access is more complex than in a data

warehouse. Individual access is possible, but one technology for individual access will be probably not sufficient. The same applies for aggregation as well. Access control can be implemented. The data governance should regulate the access to, and privacy of the data [13]. According to the characteristics, a data lake should have an information lifecycle management (ILM) strategy to manage the storage duration [13]. Since streaming data can be used and stored, it is possible to create near-real time analytics [12]. Considering that a data lake should be implemented on a scalable framework like Apache Hadoop, a data lake will be able to handle the characteristics of big data challenges [13]. Like the data warehouse a data lake is usually a repository and can be implemented locally or on a cloud platform. Many providers do offer such solutions like Google Cloud and AWS [21, 22].

5.3 Requirements from LCA

In general, it can be said that the decision criteria are dependent on the LCA use case. A company which produces clothing might fulfill criteria different than a company which produces cars or groceries. The most relevant data sources for LCA, at least for the production phase, are contained in PLM- and ERP-Systems [23]. Additionally, CAD-files [24] and descriptions of parts or orders in pure text form can also contain relevant information. So overall the LCA relevant data sources come in all structure types, which implies that a storage system for LCA use must be able to store all relevant data types as well.

The two most important data types to be ingested are application data, from PLM or ERP systems [23], and batch data like CAD-Files, tables or text files. Streaming data, like a constant stream of data from a sensor, is typically not used for LCA. An aggregation can be helpful when estimating the energy consumption of processes from sensor values, since the storage of a data stream needs a lot of memory. For the generation of an LCA report, the individual access to datasets and to aggregated data is both necessary. The necessity of access control depends highly on the LCA use case and company policies. For the purpose of a decision, it can be said that access control is not necessary by default. Usually, the LCA related data only needs to be available till the end of the LCA. To speed up future LCAs of similar products, it makes sense to store relevant data after that. On the other hand, the use of data which is too old, and not up to date, can be harmful as well.

LCA is a complex method with many aspects to consider in order to properly interpret the results and to draw reasonable conclusions [25]. For this reason, LCA can be usually seen as performance-uncritical task.

A high number of LCAs for multiple products or a LCA for a very complex product can lead to the necessity of big data solutions. One big data criterion which is relevant for LCA, is the high variation of data from different sources. But again, big data is not relevant for LCA by default. The same applies to cloud computing. The data collection and aggregation process of a product with a complex supply chain of multiple companies could benefit significantly from cloud computing. As shown in studies about the beef supply chain [10, 11]. That said, storing companies sensitive data at a third-party cloud provider could impose potential risks [26]. For this reason, there is still a lot of skepticism from companies to cloud computing.

5.4 Summary and Decision

To come to a decision, which data storage system is more suitable for the LCA use case, the hard criteria for LCA must be considered first. In the LCA use case, those are the types of structure. In general, both systems are suitable for storing data for LCA. Because the data sources of LCA are rather diverse and other use cases like LCA hotspot analysis and visualization might require the files in their raw format, a data lake would be a good choice. As shown in Table 1, a data warehouse can only store structured data, a data lake is made for storing all three types of data. All different data types can be extracted via an integration process and only structured and transformed data will be stored. Overall, as described above, the relevant data comes from a variety of sources, in all kinds of structure types and the amount can even extend to become a big data project. In all these cases, a data lake would be more suitable to store the data since it offers more flexibility than a data warehouse.

Table 1. Evaluation of data storage systems

Perspectives	Criteria	LCA (requirements)	Data Warehouse (possibilities)	Data Lake (possibilities)
Types of Structure	Structured	yes	yes	yes
	Semi-structured	(yes)	no	yes
	Unstructured	yes	no	yes
Ingest	Application data	yes	yes	yes
	Streaming data	no	no	yes
	Batch data	yes	yes	yes
Storage access	individual access	yes	simple	complex
	aggregation access	yes	simple	complex
	access controls	possible	yes	yes
	storage duration	middle	long	long
Perfor- mance	near-real time analytics	no	no	yes
Big Data	Volume	possible	no	yes
	Velocity	possible	no	yes
	Variation	yes	no	yes
	Variability	no	yes	yes
Cloud Computing	on-demand self-service	no	yes	yes
	broad network access	no	yes	yes
	resource pooling	no	yes	yes
	rapid elasticity	no	yes	yes
	measured service	no	yes	yes

6 Conclusions

Concluding, it has been found that there are many aspects to consider when choosing an appropriate data storage system for LCA. Even if both evaluated data storage systems are suitable, a data lake, thanks to its flexibility, is the better choice for the LCA use case. As a next step, a prototype of a data lake for LCA use will be implemented for further analysis. The goal is to design a data lake in a way to be appropriate for LCA. It is expected that the challenges will be the interface of the PLM and ERP systems as well as the evaluation of CAD files. Beside the data integration, there is the question how different data types will be treated and how to deal with missing information. The overall goal is to have a platform to test (semi-)automatic LCA. Other ideas like enhanced LCA with value estimation via machine learning for more exact LCA could be possible topics for further publications. In our vision, a data lake could be the base not only for the evaluation of products but could also support topics like LCA driven product design. A possible use case could be to compare different product designs. The system would get CAD geometry data and combine them with the material information of an ERP-system and the knowledge of a LCA databases to generate a CO₂-report.

Acknowledgments. The research has been applied for and was granted as COMET project under the guidance of the Austrian Research Promotion Agency FFG and is funded by the Federal Ministry for Transport, Innovation and Technology (BMVIT), the Federal Ministry for Digital and Economic Affairs (BMDW) and the provinces of Upper Austria and Styria.

References

1. He, B., Wang, J., Huang, S., Wang, Y.: Low-carbon product design for product life cycle. *J. Eng. Design* **26**, 321–339 (2015)
2. He, B., Wang, J., Deng, Z.: Cost-constrained low-carbon product design. *Int. J. Adv. Manuf. Technol.* **79**(9–12), 1821–1828 (2015). <https://doi.org/10.1007/s00170-015-6947-z>
3. Merschak, S., Hehenberger, P.: Ecodesign methods for mechatronic systems: a literature review and classification. In: 2019 20th International Conference on Research and Education in Mechatronics (REM). IEEE, Wels (2019)
4. DIN EN ISO 14040:2009-11, Environmental management - Life cycle assessment - Principles and framework (ISO 14040:2006)
5. DIN EN ISO 14044:2018-05, Environmental management - Life cycle assessment - Requirements and guidelines (ISO 14044:2006 + Amd 1:2017)
6. Crickard, P.: *Data Engineering with Python*. 1st edn. Packt Publishing (2020)
7. Sullivan, D.: *Official Google Cloud Certified Professional Data Engineer Study Guide*. Wiley, Indianapolis (2020)
8. Beier, G., Kiefer, J., Knopf, J.: Potentials of big data for corporate environmental management: a case study from the German automotive industry. *J. Ind. Ecol.* **26**, 336–349 (2020)
9. Song, M.-L., Fisher, R., Wang, J.-L., Cui, L.-B.: Environmental performance evaluation with big data: theories and methods. *Ann. Oper. Res.* **270**(1–2), 459–472 (2016). <https://doi.org/10.1007/s10479-016-2158-8>
10. Singh, A., Mishra, N., Ali, N., Shukla, S.I., Shankar, R.: Cloud computing technology: reducing carbon footprint in beef supply chain. *Int. J. Prod. Econ.* **164**, 462–471 (2015)

11. Singh, A., Kumari, S., Malekpoor, H., Mishra, N.: Big data cloud computing framework for low carbon supplier selection in the beef supply chain. *J. Clean. Prod.* **202**, 139–149 (2018)
12. Baars, H., Kemper, H.-G.: *Business Intelligence & Analytics - Grundlagen und praktische Anwendungen: Ansätze der IT-basierten Entscheidungsunterstützung*, 4th edn. Springer, Wiesbaden (2021)
13. Gupta, S., Giri, V.: *Practical Enterprise Data Lake Insights: Handle Data-Driven Challenges in an Enterprise Big Data Lake*. 1st edn. Apress®, Berkeley (2018)
14. Li, K.-C., Jiang, H., Zomaya, A. Y.: *Big data Management and Processing*. Taylor & Francis Group (2017)
15. Lakhe, B.: *Practical Hadoop Migration: How to Integrate Your RDBMS with the Hadoop Ecosystem and Re-architect Relational Applications to NoSQL*. Apress®, New York (2016)
16. NIST Big Data Public Working Group: *NIST Big Data Interoperability Framework: Volume 1, Definitions, version 2*. National Institute of Standards and Technology (2018)
17. Mell, P.M., Grance, T.: *The NIST definition of cloud computing*. National Institute of Standards and Technology, Gaithersburg (2011)
18. O’Leary, D.E.: Embedding AI and crowdsourcing in the big data lake. *IEEE Intell. Syst.* **29**(5), 70–73 (2014)
19. Data-Warehouse-Lösungen. <https://www.ibm.com/de-de/analytics/data-warehouse>. Accessed 05 Feb 2022
20. Was ist ein Data Warehouse? | Wichtige Konzepte | Amazon Web Services, <https://aws.amazon.com/de/data-warehouse/>. Accessed 05 Feb 2022
21. What is a Data Lake? Google Cloud. <https://cloud.google.com/learn/what-is-a-data-lake>. Accessed 05 Feb 2022
22. Data Lake | Implementierungen | AWS-Lösungen. <https://aws.amazon.com/de/solutions/implementations/data-lake-solution/>. Accessed 05 Feb 2022
23. Merschak, S., Hehenberger, P., Schmidt, S., Kirchberger, R.: Considerations of life cycle assessment and the estimate of carbon footprint of powertrains. *SAE Technical Papers* (2020)
24. Merschak, S., Hehenberger, P., Bachler, J., Kogler, A.: Data relevance and sources for carbon footprint calculation in powertrain production. In: Nyffenegger, F., Ríos, J., Rivest, L., Bouras, A. (eds.) *PLM 2020*. IAICT, vol. 594, pp. 203–214. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62807-9_17
25. Frischknecht, R.: *Lehrbuch der Ökobilanzierung*. Springer, Heidelberg (2020). <https://doi.org/10.1007/978-3-662-54763-2>
26. Fujita, H., Tuba, M., Sasaki, J.: *Study on advantages and disadvantages of Cloud Computing - the advantages of telemetry applications in the cloud* (2013)