A Lakehouse-based platform for Data-driven Sustainability Monitoring in Energy-Intensive Production

(Doctoral Consortium)

Paola Magrino[1],*

[1]*University of Brescia, Dept. of Information Engineering*
Via Branze 38, 25123 - Brescia (Italy)

Abstract

In energy-demanding production processes, the extraction of proper indicators from the ever-growing sources of heterogeneous data is of paramount usefulness to develop advanced data-driven applications in this context, such as anomaly detection, prediction of energy consumption, processes compliance to standards and regulations. Addressing the inherent data heterogeneity, Data Lakes has been widely promoted in recent years as a scalable, flexible, and cost-effective solution, combining them with traditional Data Warehouses or building on top of them multi-tiered architectures for the extraction of indicators from heterogeneous data sources. Such a multi-tier architecture presents several limitations, mainly due to reliability, data staleness and limited support to advanced analytics tools. Therefore, data Lakehouse technology has been introduced as a cutting-edge advancement, to offer schema enforcement, indexing, and performance optimization, while retaining the flexibility of Data Lakes, enhancing the efficiency of sustainability monitoring. In this paper, we describe the research challenges and the first steps towards the exploitation of the transformative potential of Data Lakehouses to evolve the multi-tiered architecture of an approach, named PERSEUS, designed to facilitate the personalized exploration of multi-dimensional aggregated data in a Semantic Data Lake, presenting a compelling case for the adoption of this technology in the pursuit of sustainable and environmentally responsible industrial practices.

Keywords

Data Lakehouses, Data Lakes, Sustainability indicators, Energy-intensive industrial processes

1. Introduction

In the context of energy-demanding production processes, the need for effective and comprehensive monitoring of sustainability indicators has become paramount. Sustainability can be declined according to several perspectives, namely environmental, economic and energy sustainability, and proper indicators have been developed through different standards, such as the Environmental, Social, and Corporate Governance (ESG) indicators introduced by organizations like the German Investment Professional Association (DVFA), that gained the status of official standard for the European Federation of Financial Analysts Societies (EFFAS) [1], or the set of standards developed by the Global Reporting Initiative (GRI) [2], or other frameworks and standards that are underway towards homogenization and harmonization [3]. Indicators are

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

*Corresponding author.

paola.magrino@unibs.it (P. Magrino)

extracted from largely available heterogeneous data sources to develop advanced data-driven applications, such as anomaly detection, energy consumption prediction, compliance of processes to standards and regulations. Data Lakes have been proposed in recent years as a novel and efficient solution to address the challenges posed by the heterogeneity of data generated in such complex environments. Data Lakes, with their ability to store and process diverse data types and formats, offer a promising framework to aggregate and analyze the multitude of information sources related to sustainability metrics as well. Highlighting the positive characteristics of Data Lakes, we emphasize their scalability, flexibility, and cost-effectiveness. These attributes enable organizations to seamlessly integrate large volumes of data from disparate sources, fostering a holistic approach to sustainability monitoring. Additionally, the ability to handle unstructured and semi-structured data enhances the inclusivity of the analysis, capturing valuable insights that may be overlooked in more rigid data storage architectures, such as the ones of traditional Data Warehouses. In recent years, several approaches have been proposed to extract indicators from Data Lakes [4, 5, 6, 7]. Among them, we mention PERSEUS (PERSonalised Exploration by User Support), a three-phase computer-aided approach facilitating personalized exploration of multi-dimensional aggregated data in a Semantic Data Lake, involving the construction of a semantic metadata catalog, modeling indicators and analysis dimensions using a Multi-Dimensional Ontology, and enriching indicators with personalization aspects based on users' profiles and preferences to enable interactive exploration.

Nonetheless, although the schema-on-read architecture of the Data Lake allowed for flexible storage of various data types at a low cost, it deferred the challenges of ensuring data quality and governance to a later stage. In this framework, a portion of the data within the lake would undergo ETL processes for advanced data management, specifically for critical decision support and business intelligence applications. This results in the design of a multi-tier architecture, where Data Lakes and Data Warehouses are used together, or to the multi-tiered architecture of PERSEUS, where data from heterogeneous sources undergoes a complex set of steps, including semantic layer construction and knowledge graphs extraction for their personalised exploration. Such a multi-tier architecture presents several limitations, mainly due to reliability (it is difficult and costly to keep the Data Lake and upmost layers consistent), data staleness (the data in the upper layers is stale compared to that of the Data Lake, as new data coming from IoT systems is frequently loaded) and limited support to advanced analytics tools such as TensorFlow, PyTorch and XGBoost, that offer advanced solutions for predictions over the status of sustainability indicators.

To improve the efficiency and capability of sustainability monitoring, the cutting-edge technology of Data Lakehouses has been introduced [8, 9, 10]. Data Lakehouses combine the best features of Data Lakes and traditional Data Warehouses, offering the advantages of schema enforcement, indexing, and performance optimization, while retaining the flexibility and scalability inherent to Data Lakes. This innovative approach ensures a streamlined and powerful analytical environment, facilitating real-time decision-making for sustainability initiatives.

The aim of this paper is to describe the research challenges and the first steps towards the exploitation of the transformative potential of Data Lakehouses to evolve the architecture of approaches such as the PERSEUS one, discussing the possible advantages for the adoption of this technology in the pursuit of sustainable and environmentally responsible industrial practices. Unlike traditional data, sustainability data requires specific considerations due to its

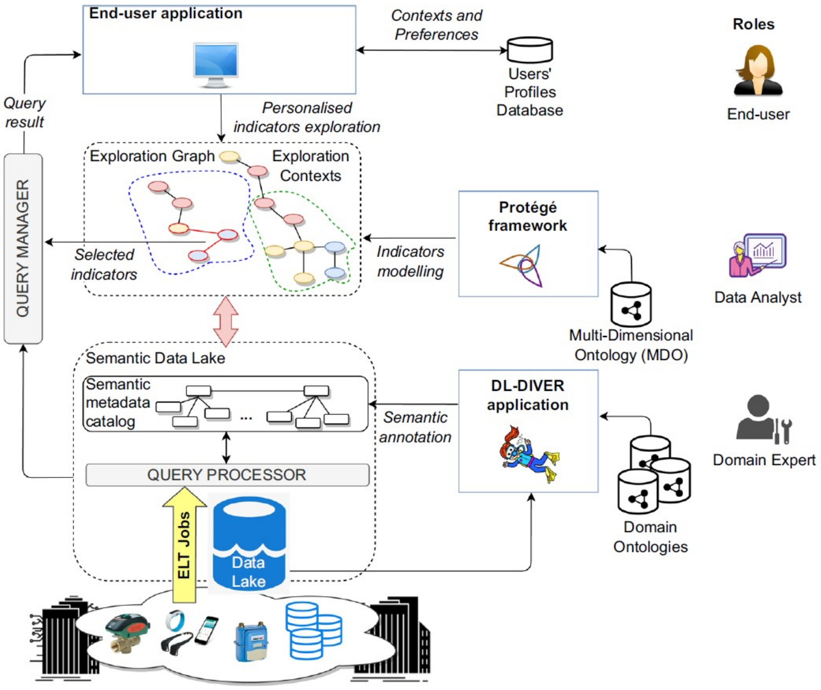

Figure 1: Architecture of the PERSEUS (PERSonalised Exploration by User Support) approach

qualitative and quantitative nature, diverse sources, and the necessity for accurate reporting on ESG performance in terms of environment, social and governance factors. These unique characteristics, including complexity and diversity, regulatory compliance, stakeholder engagement, a long-term perspective, and interconnectedness of sustainability issues, allow to build an architecture that adapts to those specific requirements. However, this architecture can also be generalized to other contexts where similar requirements remain relevant, enabling broader applicability.

The paper is organised as follows: in Section 2 the PERSEUS approach is briefly described; Section 3 provides an overview of the proposed Lakehouse-based architecture for data-driven sustainability; future research challenges are discussed in Section 4; finally, Section 5 closes the paper.

2. The PERSEUS approach

The PERSEUS (PERSonalised Exploration by User Support) approach (Figure 1) is a computer-aided approach, articulated over three phases, to build Exploration Contexts on top of the Semantic Data Lake [11]. The phases of the PERSEUS approach are conceived to enable personalised exploration of multi-dimensional aggregated data, by means of indicators modelled on top of Data Lake sources, by progressively enriching the organisation of Data Lake content in a (semi-)automatic way through:

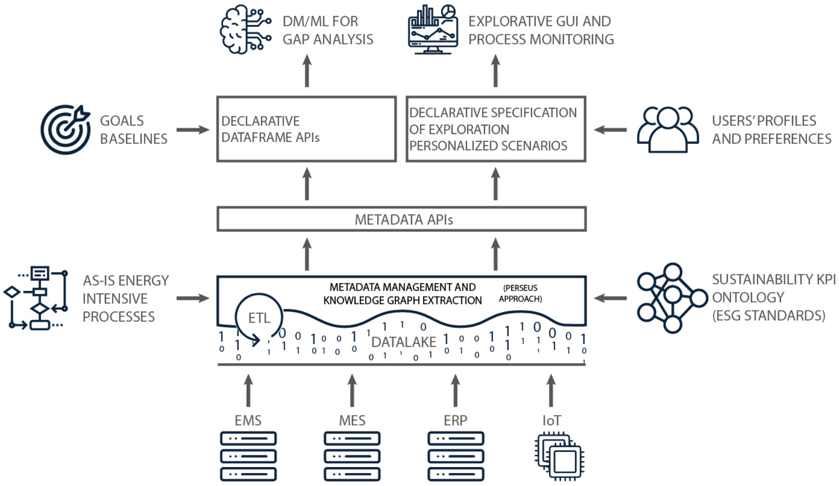
DM/ML FOR GAP ANALYSIS
EXPLORATIVE GUI AND PROCESS MONITORING
GOALS BASELINES
DECLARATIVE DATAFRAME APIs
DECLARATIVE SPECIFICATION OF EXPLORATION PERSONALIZED SCENARIOS
USERS' PROFILES AND PREFERENCES
METADATA APIs
AS-IS ENERGY INTENSIVE PROCESSES
METADATA MANAGEMENT AND KNOWLEDGE GRAPH EXTRACTION
(PERSEUS APPROACH)
ETL
DATALAKE
SUSTAINABILITY KPI ONTOLOGY (ESG STANDARDS)
EMS
MES
ERP
IoT

Figure 2: Proposed architecture of the Lakehouse-based platform for data-driven sustainability.

- the construction of a semantic metadata catalog on top of the Data Lake, to semantically describe sources metadata; proper tools and metrics have been defined to support the annotation process of the metadata catalog; the Data Lake equipped with the semantic metadata catalog is denoted as Semantic Data Lake;
- modelling of indicators and analysis dimensions, based on metadata contained in the catalog, leveraged by users to explore Data Lake contents as aggregated multi-dimensional data; semantic representation of indicators and analysis dimensions is denoted as Exploration Graph; semantic modelling is guided by a Multi-Dimensional Ontology, that is openly available and explorable through well-known tools (e.g. Protégé), and can be used to perform automatic conformance checking over the defined Exploration Graph;
- enrichment of the definition of indicators with personalisation aspects, based on users' profiles and preferences, to identify different Exploration Contexts within the Exploration Graph, for enabling personalised and interactive exploration of indicators; Exploration Contexts make easier and more usable the exploration of the data in the Data Lake for a large number of users.

3. Lakehouse-based architecture overview

The proposed architecture is based on the PERSEUS approach, using a Data Lakehouse structure to overcome its limitations (Figure 2). The architecture draws from heterogeneous data sources, including structured sources like EMS (Energy Management Systems), ERP, and MES (Manufacturing Execution Systems), as well as unstructured data from IoT (environmental sensors or machine sensors). All this information flows into the Data Lake, where the Extract, Load, Transform (ELT) process is continuously engaged for metadata management and knowledge graph extraction, thus implementing the lakehouse vision [8] as an extension of the PERSEUS approach.

An ontology will be used to describe Key Performance Indicators (KPIs). In this case, the focus will be on sustainability KPIs, organized within a reference standard. Additionally, modeling of AS-IS industrial processes is considered to relate data extracted from heterogeneous sources with energy-demanding processes and to calculate the indicators in the scope of industrial context in which data is originated, properly segmenting it with respect to the stages of the industrial production. The industrial processes will be modeled according to the three scope categories used to classify an organization's greenhouse gas emissions.[1] However, the integration of process-related information with enterprise or sensor data is an open research issue that requires further investigation. On top of the Lakehouse metadata layer lies a metadata API, that supplies data to two distinct modules, one for supporting data analytics through state-of-the-art Machine Learning models, and the other one for supporting users in the exploration of indicators, as promoted in the original PERSEUS approach by means of personalised exploration scenarios.

The first module entails a declarative dataframe API which, combined with the definition of company or institutional goals and baselines, provides data in the necessary format for data mining and machine learning algorithms to perform gap analysis and consumption forecasts. Operating directly on dataframes ensures a significant efficiency advantage over traditional approaches based on data warehouses and relational databases, which require additional steps to combine data from different sources/tables. The second module will consist of a declarative specification of PERSEUS personalized exploration scenarios. In this phase, leveraging users' profiles and sustainability-driven preferences, the necessary information for targeted exploration of indicators through a graphical interface (GUI) is obtained, facilitating the identification of any anomalies or potential inefficiencies in sustainability-aware industrial processes.

4. Research challenges

Ontologies for sustainability As already underlined in [3], one of the recent research challenges is the absence of a shared definition of metrics for sustainability. The analysis of standards reveals format, structural and terminological heterogeneity to define ESG indicators. Semantic technologies are proposed as a solution to standardization purposes. Knowledge Graphs are proposed in [3] for modeling ESG indicators, requiring interpretation of underspecified definitions. The intrinsic flexibility and modularity of a graph structure make it easier to add, modify, and reason about ESG indicators incrementally, adapting to continuously evolving industrial requirements and standards. Using an ontology-based approach to guide information extraction from industrial processes provides a formal structure and explicit meanings and relationships between concepts, including mathematical relationships and dependencies between ESG indicators. Following the PERSEUS approach, ontologies can be engaged to guide the extraction of knowledge in the data Lakehouse, but how to extract "on-the-fly" and efficiently the knowledge from data sources and manage mappings between extracted data and target industrial processes are still under-investigated tasks. Moreover, other concepts and semantic relationships must be included in the ontology, beyond indicators and dimensions,

[1] Scope 1 includes direct emissions from sources owned or controlled by the organization. Scope 2 concerns indirect emissions associated with purchased energy. Scope 3 encompasses all other indirect emissions along the value chain, such as supplier transportation and waste disposal.

such as compliance with regulations and organizational policies in energy-intensive production environments and relationship between indicators and target users, to enable personalised and sustainability-driven preference-based exploration of the extracted knowledge.

Metadata layer definition The metadata layer constitutes a fundamental aspect for data management in a Lakehouse environment, encompassing challenges that are unique to this hybrid architecture. It serves as a natural point to implement data quality control and governance features, such as applying constraints on the data schema, access control, and ensuring compliance with regulations and organizational policies [12]. Metadata layers, like Delta Lake, also confront challenges in metadata cataloging and discovery, as well as in capturing and tracking metadata lineage and provenance to provide visibility into the origin and transformations of data [8]. Within the scope of this paper, the main research objective concerning the metadata layer is about the implementation of the PERSEUS stages directly within the Lakehouse. This will require to address metadata governance issues, such as applying constraints on the data schema and computer-aided solutions for data quality control and governance, ensuring ACID transactions and other data management functionalities, and investigating how to make these stages as more automatic as possible, for example with the support of LLMs [13].

Declarative specification of exploration scenarios. Declarative query languages pose significant challenges within data lakehouse environments. They must effectively manage and integrate with metadata systems to ensure comprehensive metadata support in the Lakehouse, addressing factors such as data lineage, provenance, and governance. The focus in literature is mainly on balancing performances of query execution time versus resource utilization, to meet the needs of different applications and users. Addressing these challenges requires research and development efforts to advance the capabilities of declarative query languages within data lakehouse. In the context of PERSEUS approach for personalized exploration, research should investigate the declarative formulation of exploration scenarios to enhance their performance and portability, abstracting from implementation details. Additionally, translation of personalized exploration scenarios into SQL query plans is a significant challenge to address. Finally, designing user-friendly query interfaces and tools is essential for enabling users with varying levels of expertise to interact effectively with data lakehouse environments.

5. Conclusions

To address effective and comprehensive monitoring of sustainability indicators in energy-demanding production processes, this paper introduces the concept of Data Lakehouses, which combine the best features of Data Lakes and traditional Data Warehouses. The paper proposes a Lakehouse-based architecture for data-driven sustainability, to evolve the multi-tiered architecture of an approach, named PERSEUS, designed to facilitate the personalized exploration of multi-dimensional aggregated data in a Semantic Data Lake, presenting a compelling case for the adoption of this technology in the pursuit of sustainable and environmentally responsible industrial practices. An overview of the proposed architecture is introduced, but several research challenges remain to be investigated. These include the standardization of sustainability metrics using ontologies, the definition of metadata layers for effective data management, and the development of declarative specification for exploration scenarios. Addressing these challenges will be crucial for unlocking the transformative potential of Data Lakehouses in advancing sustainable and environmentally responsible industrial practices.

References

[1] EFFAS - The European Federation of Financial Analysis Societies, KPIs for ESG: A Guideline for the Integration of ESG into Financial Analysis and Corporate Valuation, 2009. URL: https://ec.europa.eu/docsroom/documents/1547.

[2] Global Sustainability Standards Board, Consolidated Set of the GRI Standards 2021, 2022. URL: https://www.globalreporting.org/standards/download-the-standards.

[3] C. Diamantini, T. Khan, D. Potena, E. Storti, Shared Metrics of Sustainability: a Knowledge Graph Approach, in: Proc. of the 30th Italian Symposium on Advanced Database Systems, SEBD 2022, Tirrenia (PI), Italy, June 19-22, 2022, volume 3194 of *CEUR Workshop Proceedings*, 2022, pp. 244–255.

[4] A. Pomp, A. Paulus, A. Kirmse, V. Kraus, T. Meisen, Applying Semantics to Reduce the Time to Analytics within Complex Heterogeneous Infrastructures, Technologies 6 (2018) 86.

[5] M. Pingos, A. S. Andreou, A Data Lake Metadata Enrichment Mechanism via Semantic Blueprints, in: Proc. of the 17th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2022), 2022, pp. 186–196.

[6] Q. Yuan, Y. Yuan, Z. Wen, H. Wang, C. Chen, G. Wang, Exploring Heterogeneous Data Lake based on Unified Canonical Graphs, in: Proc. of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1834–1838.

[7] D. Sarramia, A. Claude, F. Ogereau, J. Mezhoud, G. Mailhot, CEBA: a Data Lake for Data Sharing and Environmental Monitoring, Sensors 22 (2022).

[8] M. Armbrust, A. Ghodsi, R. Xin, M. Zaharia, Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics, in: Proc. of the 11th Annual Conference on Innovative Data Systems Research (CIDR 2021), 2021.

[9] A. Harby, F. Zulkernine, From Data Warehouse to Lakehouse: A Comparative Review, in: Proc. of the 2022 IEEE International Conference on Big Data, 2022, pp. 389–395.

[10] J. Schneider, C. Groger, A. Lutsch, H. Schwarz, B. Mitschang, Assessing the Lakehouse: Analysis, Requirements and Definition, in: Proc. of the 25th International Conference on Enterprise Information Systems (ICEIS 2023), 2023, pp. 44–56.

[11] D. Bianchini, V. D. Antonellis, M. Garda, A Semantics-enabled Approach for Personalised Data Lake Exploration, Knowledge and Information Systems 66 (2024) pp. 1469–1502.

[12] D. Mazumdar, J. Hughes, J. Onofré, The Data Lakehouse: Data Warehousing and More, arXiv (2023). arXiv:2310.08697.

[13] M. Nasseri, P. Brandtner, R. Zimmermann, T. Falatouri, F. Darbanian, T. Obinwanne, Applications of Large Language Models (LLMs) in Business Analytics – Exemplary Use Cases in Data Preparation Tasks, in: Proc. of the 25th International Conference on Human-Computer Interaction (HCI 2023), 2023, pp. 182–198.