Open Access to Big Energy Data Using Data Lakes

Research Essay

DLMDSSCTDS01 Current Topics in Data Science

M.Sc. Data Science, 92014910

Aaron K. Althauser

IU International University of Applied Sciences

Frankfurter Allee 73a 10247 Berlin, DE

June 24th, 2022

Contents

Open Access to Big Energy Data Using Data Lakes

1 Introduction

There is a modern notion due to the advancements in data science, that if we have enough data to tell the whole story, there is less of a need for expertise and more of a need for sufficient datasets. This rings true in the era we are living in, where Big Data is becoming more of a lucrative and replenishing resource than fossil fuels. The question on many experts' minds is how to control this massive flow of data and transform it from raw data into extracted knowledge. Big Data's five V's (Variety, Volume, Value, Veracity, and Velocity) exemplify the obstacles that data scientists face. Variety is manageable when data originating from a copious number of sources can still be linked together in the same system; a fundamental perspective of data science [1]. Volume is confounding: 2.5 Exabytes of data was the presumed global data production amount in 2012 alone [2]. Velocity is important concerning the detail that it takes about three hours to read one terabyte of data at 100 megabytes per second (mbps) speed [3]. Veracity and Variety are increasingly difficult as the heterogeneity of data sources widens. Value is only extracted if data is properly processed and maintained.

Both public and private companies in the year 2022 are still relying on outdated storage and delivery formats, i.e., storing data in spreadsheets and sending to colleagues via cloud uploaded zip files [4]. Meanwhile, researchers claim that it's no longer efficient to store data in separate locations before moving it to special computing facilities for analysis [5], as consolidating data into collections of unmanaged data would increase the reuse and sharing capabilities and would allow enterprises to cut costs by reducing both servers and licenses [6]. Big Data architecture is an obvious solution to this; but how do we blend massive volumes of data from heterogenous sources while trading redundancy for complexity?

Lansing et al. (2021) contend that the adoption of open-source architecture and 'shared solutions to common data needs' will not only reduce the effort it takes to duplicate findings but will cut down the time it takes to transfer analyses into commercial results. Shared solutions can be in the form of open access data portals that allow for unlimited distribution and transformation. Open access supports the idea that data should be in reusable formats and freely accessible to drive innovation [8].

Relying on proprietary data services is not sustainable in the scientific community, where researchers commonly rely on grants and public funding. As well, the coordination between private and public sector stakeholders does not inspire innovation. Giest (2017) highlighted a lack of harmonization between governments and private sector energy providers in Vienna, Copenhagen, and Oxford, where the city municipalities utilize Open Data portals to collate disparate datasets from smart city sensors for urban innovations and emission reduction campaigns. Similar innovations need to be discussed in parallel sectors, i.e., the building sector, where research has been conducted to identify the practicalities of organizing a large-scale database to monitor building performances from around 750,000 commercial and residential buildings. The collection of datasets comes from data from cities, building owners, and energy efficiency programs, and concludes that there is a demand for Big Data in the building sector [3]. Similarly, Mathew et al. (2015) agree there is no uniform format for data collection and storage for the building sector to access energy-related data.

In the renewable energy sector, researchers are calling for an industry standard in calculating Annual Energy Production (AEP) of renewable energy plants, as industry experts must currently rely on commercial tools and data acquisition providers [10]. As hardware is becoming cheaper [11], there are more prime opportunities to mine expansive volumes of data to integrate renewable energy into more accessible locations. Data used for Big Energy, as with all other sectors that rely on its resource, needs to be integrated and mutated into correct formats for companies to utilize the data [12], and therefore specific data structures that are scalable to house extreme volumes, optimized to provide practically real-time analyses, and cost-efficient to be considered by both profit-based and publicly funded institutions, need to be designed.

One modern Big Data architecture that is increasingly gaining attention in the scientific community the Data Lake (DL). DL is a blanket term for a methodology of storing, archiving, refining, and exploring data within a prodigious repository of data that is based on low-cost, distributed, and horizontally scalable architecture [6]. DL are starting to see more use for sectors that depend on copious volumes of data: Sarramia et al. (2022) propose an open DL for the storage and distribution of heterogeneous environmental data for the use of research programs and environmental studies, while other research in the renewable energy [24] [10] and building sectors [3] [5].

The scope of this report will begin with a high-level overview of DL and their imbedded layers and technologies, and how enterprises can design DL to attend to modern problems. Along with exploring the best architectural paradigms needed to follow suitable data security and operability measures, it will compare the use of DL with other Big Data infrastructures and examine the lifecycle of data stored in DL, along with the challenges involved at each step. The report will then cover renewable Big Energy and Big Climate data, respectively, from the perspective of using DL and Open Data portals to both proliferate public and private stakeholder knowledge in sustainable areas and drive innovation. Finally, further research in using DL and related technologies will be presented to promote open access to data for the renewable energy sector.

2 Data Lakes

Data Lakes provide capabilities to extract knowledge from Big Data amounts of semi-structured raw data using data mining techniques at the ingestion, storage, and exploration stages; all while enabling the use of modern data technologies at an affordable cost [13]. Due to the need to scale horizontally, where increase in scale is mitigated by low-cost, distributed machines, typically in cloud networks—as opposed to a vertical scaling where more processing power means the need for larger machines with more RAM—DL are commonly built on cloud infrastructures like Amazon Web Services (AWS), Azure Data Lake Store, and Google Cloud Platform (GCP) [14]. They are commonly built on top of NoSQL databases. NoSQL refers to frameworks that use a Database Management System (DBMS) not confined to relational data structures and strict schema languages, as opposed to earlier SQL-based DBMS which only allow structured data to be stored and accessed. Databases such as MongoDB, CouchDB, Dynamo, and Cassandra are progressively becoming mainstream NoSQL databases [11]; the main advantages being they are open-source and cloud-based technologies designed to store large amounts of information. These tools are fundamental building blocks for DL.

2.1 Architecture

Data lakes contain four main layers of operation: Ingestion, Storage, Transformation, and Interaction [15]. The ingestion layer absorbs data into the DL from multiple sources and extracts metadata for later access. The storage layer provides long and short-term storage of data in its original and raw form using a centralized collection of repositories. The transformation layer handles user queries and access to the data. The interaction layer involves all processes of filtering for data visualization and analytics.

For data to be usable, DL process raw data and discover metadata within the data, to find relationships unknown before, and to enact proper Extract-Transfer-Load (ETL/ELT) processes. Metadata extraction is important for the lifecycle of data residing in DL, as schema may change over time or be unknown. The different types of metadata that exist are schema preserving dataset structure, semantic metadata and constraints, and descriptive information. The flow from raw data to quality data starts with the extraction of schema, semantics, and ingested data; then formal metadata modeling takes place for organization; enrichment of the metadata is then possible, i.e. to identify relatedness with other datasets; then after schema mappings and/or human annotations are created, the metadata can enable queries of the DL and data quality improvements. Frameworks such as the Generic and Extensible Metadata Management System (GEMMS) operate as a metadata controller by forming metamodels from extracted metadata of heterogeneous sources [14].

2.2 Data Lifecycle & Governance

Data governance plays a vital role in the quality of information one can extract from DL. It is critical to monitor data quality using both qualitative and quantitative methods to decide if data is either destructive or valuable, because data at any given time will need to move about the DL. The ability of raw data to move is dependent on constraints set by data governance principles such as compliance, security, sensitivity, and data volume limits [16]. This has an influence on the design and implementation of DL by restricting the data to the confines of an organization's IT infrastructure and storage capabilities.

To avoid the implications of DL data becoming convoluted, otherwise known as a 'data swamp', organizing metadata at the precise moment of ingestion into the DL is necessary for organizational purposes [14], while governing how data evolves as its metadata is constantly re-written is fundamental to the DL archetype. Data swamps are inevitable when the lifecycle of data entering the lake is not accounted for, and when proper data governance is not abided by for data retention and handling [13].

Databases change over time, especially when governed with flexible schema. Inserting, deleting, and modifying operations of metadata and schema change the structure of the original, raw data; therefore, continued use of the same data inevitably leads to misleading results [28]. Because operational business databases are constantly mutating in size and form to answer analytical queries, they should be dynamic and capable of utilizing any other resource to reserve computational space and time. Researchers Derakhshannia et al. (2020) use paradigms from ecological models of natural lakes to conceptualize DL and the data lifecycle, and use comparisons between concepts to propose a more sustainable data governance policy. Researches similar to these are beneficial to innovating DL technology as governance of data is one of the main challenges faced and one that similar data structures can overcome.

2.3 Data Warehouses

Data Warehouse (DWH) are an enterprise structure meant to house data from various sources into 'one source of truth'; A source of analytics that the organization could access at any time, repetitively and consistently [6]. DWH are used more in practice by businesses that require quick analytical processing, but there are quite a few distinctions between the different architectures. There is a distinction between the expertise of who is accessing the data: DWH are suited better for end-users because they are information-driven, while DL are more accessible by 'power-users' like analysts and scientists [16]; when controlling ETL/ELT processes, DWH systems are designed with strict governing policies for importing data, and highly organized for making queries, while DL contain a wide scope of data structures and types [17]; DWH are built with rigid and complex structures that are time-consuming to change, while metadata in DL can be re-written relatively infinitely; DWH only contain structured data, while DL contain both semi-structured and structured; DWH data is mainly stored for reporting purposes, while DL store all data comprehensively in case different analyses need to be supported.

DL are, in short, a methodology of combining all data sources within an organization into a central, accessible location. When data scientists extract data from the DL, the data then populates a DWH that is more structured for query processing [16]. This is like what Fang (2015) pointed out, that the hybrid use of both DL and enterprise DWH will be the necessary data ecosystem in place in the future, where users could ask real questions and have their questions backed up by sufficient data. Likewise, the industry is beginning to be dominated by the federated DL and warehouse technologies used in tandem. In such systems, researchers argue the inevitability of the sources of data and to some extent the analytics being separated and unknown from the end users [18].

2.4 Modern Use Cases

DL are starting to be developed for many different sectors and purposes. Deligiannis et al. (2020) developed an online DL for the distribution and reuse of heterogeneous cultural heritage data using open-source tools, presenting a tool called Hydria that allows non-IT users to access, analyze, correlate, and visualize data from disparate sources and formats. Lansing et al. (2021) developed the Time Series Data Pipelines (Tsdat) to support open-source marine energy data, with the purpose of exposing marine data for its analysis through a structured DL. Researchers in another study propose a personal DL architecture, where individuals' information given to data providers is collected in individual DL [20]. Data access is then controlled by individuals by allowing or disallowing access to certain datasets within their individual lake. Such use cases of DL require more thorough research when it comes to overcoming some common challenges in deploymentL.

2.5 Challenges

Schema re-writes. Lazy integration, where repositories are mainly kept in their original form and linked to DL through schema-mapping and re-writing, remains one of the most difficult obstacles faced by researchers designing DL systems.

Data mining. Machine learning tools are traditionally designed for the extraction of knowledge only from raw data, which is not consistent with modern big data systems that have multiple data types, so the challenge of how to intelligently extract information remains for data scientists [6]. It is also a problem in DL for analysts to understand the data quality of past analyses by others who have accessed the data and restructured some relationships between datasets.

Security. Often highly sensitive data is stored in DL, and when access to the data is not protected by transactional queries, the level of clearance needed to access sensitive data can occasionally become avoidable. Relational Database Management System (RDBMS) allow access to databases using ACID principles (Atomic, Consistent, Isolated, and Durable).[1]

Performance. As Fang (2015) states, “Tools and data interfaces simply cannot perform as

[1] https://www.ibm.com/docs/en/cics-ts/5.4?topic=processing-acid-properties-transactions

well against a general-purpose store as they can against optimized and purpose-built infrastructure.”, which highlights that future developments for DL should be focused on how to compete performance-wise with other Big Data systems.

3 Open Data

Data being termed ‘Open Access’ or ‘Open Data’ represents a movement where individuals and groups are able to freely access data that provides social and environmental benefits, among others. The idea is that the value of datasets is multiplied when added all together in a ‘playground’ of data, leading to competitive and, in turn, innovative data. For data to drive innovation, it is argued that it must first be in reusable formats and freely accessible [8]. This approach should be applied to Big Climate and Big Energy data to support the innovation of technology that provides livelihood to communities. To guarantee environment and energy data are reproducible, interoperable, reusable, and accessible, it is important to follow the FAIR principles (Findability, Accessibility, Interoperability, and Reusability) when implementing strategies for organizations [21]. OpenDataMonitor [2] is a great example of a platform that provides access to both national and international Open Data portals to end-users, while providing advanced search and visualization features. Another example, HackAir [3] is a European platform that monitors air quality through Open Data and provides realtime forecasts that are customizable to specific needs.

For data to be considered open, it must be attributed and shared non-commercially. It must also be accessible by being offered at a fair cost, preferably over the internet, and provided in its entirety. Licensing terms of the data should allow it to be reused and redistributed and should allow any and all modifications to take place. Finally, Open Data should be completely objective and void of any discrimination against individuals or groups [22]. In one study [7], it was highlighted that while many end-to-end data integration pipeline solutions exist in the commercial market (Talend, Tableau, Domo) that evidently provide much value to their users, their licensing costs are not affordable for scientists and startups needing to use the services. Additionally, their closed-source nature alludes that making customizations or sharing components across projects is not supported.

Besides the overwhelming economic benefits of Open Data, openly sharing data leads to more efficient processing, standardization, and automation; but it also has a massive global impact in the form of environmental benefits [8]. By opening the large accumulation of statistical data on CO2 emission, Open Data provides more insight and a better understanding of problem factors for those in positions to make actionable policy changes. For example, in Austria, data on household energy consumption is held by both private energy providers and the City of Vienna using Open Data portals; but there is a problem in architecture trying to link the two together, resulting in data not being used to

[2] https://www.opendatamonitor.eu/

[3] https://platform.hackair.eu/

its fullest potential [9]. One report promotes the idea of integrating data from the energy sector into the transport master plan, bridging the gap between public utilities and experts in the climate and energy sectors [9], while researchers involved in building energy use are constructing a national database to increase accessibility to energy data for the building sector [5].

Strategies for the opening data for use of the public are starting to receive funding, as they are increasing the economic and social benefits of societies. The city of Copenhagen was partnering with Hitachi to create a ‘City Data Exchange Hub’, a central storage location for data intended to be bought by businesses or accessed by the public to innovate green infrastructure planning and optimize energy efficiency [9].

3.1 Policies

At the intersection of Big Data and sustainability, there is a modern practice of governing bodies using the data movement to optimize public areas and redefine civic engagement while offering smart solutions for urban spaces. This is loosely defined as ’policy-making 2.0’ and provides a platform for evidence-based policy-making and governance [9]. Open Data for public administrations was valued at 22 billion EUR in 2020 [23], showing how the public sector has economic reasons for being the first to reuse Open Data.

Data-driven decision and policy-making for governing societies depends on access to large volumes of both historical and current data [17]. CBECS and RECS—-two U.S. national statistical datasets, are relied on for the U.S. Energy Information Administration’s (EIA) Annual Energy Outlook [5]. Reuse of Open Data through portals has had other inspiring developments for societal benefits: OpenCorporates helps organizations learn about other companies’ environmental impact, financial stability, and networks through an open database [4]; CityScale provides Open Data on crime rates, air pollution and healthcare for Ukrainian citizens [5]; Plantwise provides information about disease threats and health of plants for farmers through the collection of Open Data [6]; and CIARD is a central repository of agricultural information and research with more than 1,500 open agricultural datasets of research. [7]

On the other hand, it has been shown that less companies were reusing earth observational data due to the lack of data being shared on national data portals, unlike company or transportation data [8]. Regulations surrounding the competitive markets of electricity are being written by economic factors such as gas and oil prices, and market stability. The flow of market information going into data control centers only adds to the ‘data tsunami’ that data scientists need to control [11].

[4] https://opencorporates.com/info/about

[5] https://www.cityscale.com.ua/

[6] https://www.plantwise.org/

[7] http://www.ciard.net/

Local governments play an ever-increasing role in the action taken to resolve global issues in the areas of environmental and social sustainability. They typically stick to two strategies to make use of Big Data. In one, they utilize cloud computing to integrate all data into one database, creating a centralized storage space for all users to access (in one local data center); in the other, they use a distributed model to collect pieces of data from various sources, dispersing data scientists across different regions and departments. With the goal of making Copenhagen carbon neutral by 2025, the city municipality adopted the CPH 2025 Climate Plan. It is an action by the government to use Big Data as a tool to reduce energy consumption. The city is also working with IBM to create an 'Open Data Hub'—another way of fighting carbon emissions by using Big Data—the hub connects consumers, providers, entrepreneurs, programmers, and any interested parties to challenge and motivate them to create solutions based off Big Data analytics, with the goal to decrease the city's overall energy use [9]. The process involves a multitude of advanced sensors to monitor and collect data and directly turn data into policy initiatives.

3.2 Challenges

Politics. Using Big Data in an evidence-based policy-making environment challenges the archaic methods of improving societies. Although the advantages of using Big Data for complex analytics are largely due to being able to extract knowledgeable insights directly from the data, politicians rely on monolithic initiatives and legislature that require majority votes. The passing of legislature is not directly influenced by the amount of data present when making the decisions, but instead the public feedback and personal gains, and therefore policy-makers do not have the proper motive to provide open access to data.

Private sector. Market leaders in private corporations rely on data science to propel the businesses beyond profit margins. The same leaders are the ones who have the most access to usable data, especially in the energy sector. Similarly, stakeholders representing private data providers depend on proprietary solutions using data as a resource.

4 Big Data and Renewable Energy

Big Data and the renewable energy field are symbiotic in relationship. The challenges of globally reducing carbon emissions are inexorably linked with efficiency and performance of Big Data systems. Reducing carbon emissions, for example, involves a wide inventory of different data sources and types, making it a necessity to use optimized and efficient architecture during analysis. Analytics of systems intended to house Big Energy data will provide decision makers with sufficient evidence to push society forward in a more sustainable direction. Currently, there is an issue of how to integrate data between the host of sensors in renewable energy plants. Smart grids rely on a network of sensors and multitudes of smart meters all talking to each other. Standardization is one way to solve the data integration issue commonly faced with Big Energy data [12]. Standardizing data is a way to facilitate the communication between the network of sensors, to allow better data exchange and interconnection

There is a deficiency of current research into Big Data systems that are specifically designed to manage and analyze renewable energy production sites [24]. In a recent study, OpenOA was created to provide operational analysis to the renewable energies sector, particularly to the analysis of wind farms, by calculating AEP, electricity and turbine level optimization [10]. Ceci et al. (2014) present a prototype for managing renewable energy production plants using a column-store database (HBase) to separate plant information, plant measurement, and prediction data. What the prototype does not offer is the integration of varied data sources into its architecture, as it is operationally expensive for HBase to use more than three columns to store data.

4.1 Designing A Big Data System

Renewable energy sources are commonly controlled by a grid technology that distributes and transmits varying levels of capacity throughout the grid at calculated intervals [11]. Operators have begun to take advantage of smart grid technology advanced enough to predict when optimal resources can be utilized from each unit in the grid, i.e. when regional weather conditions will be fruitful for harvesting power. To balance both consumption and production, it is crucial that smart grids are not consuming too many resources in non-optimal periods, which can be controlled by the fact that smart grids generate massive volumes of information on an hourly basis. How this amount of information is properly handled determines how useful the information is to the market.

To maximize efficiency for a renewable energy system, there needs to be responsible tradeoff between storage and analysis of data [24]. For Big Data systems to efficiently handle modern analytical tasks, there needs to be guidelines that follow these general rules: a wide variety of analytical tools need to be supported, data should be accessible using open-source tools and retrieved with API services, the system should be able to process streaming data without being overwhelmed, and it should provide a high level of security of the processed and stored data [24].

4.2 Challenges

Marketability. Because development of Big Data systems for renewable energy purposes requires vast knowledge of a plethora of tools and expertise, they typically only can be developed by big players in the market who have primarily financial motivations for building such systems [24]. Therefore, the technology needs to be developed in ways that allow for low-cost, low-expertise analysis by end-users.

Hardware. Power grid systems are unmanageable when there are too many inconsistencies in the data. While it is nearly impossible to eliminate all inconsistencies, it is possible to develop a more consistent power grid system, that owners in the market can use to efficiently forecast and predict power usages.

Because renewable energy plants are highly dependent on local weather patterns, it is important to increase the amount of Open Data generated from environment data. Using shared DL, scientists would be able to access data not just from regional power plants, but they could harvest all incoming environment data to fully optimize production of renewable energies.

5 Big Climate Data

Advances in technology have increased the amount of collected data on the environment. While most of this data is collected for projects with a specific focus, much of the data could be reused for cross-analysis between different departments; but merging data from different projects is always wrought with incompatibilities and inconsistencies in the collected data, which calls to light the need for more flexible and powerful storage systems to integrate and analyze the data [9].

Attaining knowledge of a specific region's potential to provide sustainable energy requires vast knowledge of the intricacies of the territory [21]. To collect said data, it is necessary to draw from all available sources: the United Nations Framework Convention on Climate Change (UNFCC) highlights 34 variables of climate data that require collection from satellites [25]; Researchers Faghmous and Kumar (2014) highlight an idea of a "climate network" that utilizes graph theory to place weights between nodes of geographic location, with weights depending on the time series characterization of each node; and a collection of over 100 climate datasets, including peer-reviewed publications that used the data, are accessible through the community-created Climate Data Guide sponsored by the U.S. National Science Foundation. [8]

5.1 Challenges

Data lifecycle. Climate scientists dealing with Big Data are challenged when it comes to governance and federation, along with distribution and archival of massive volumes of data [26].

Hadoop. While climate scientists mainly deal with spatiotemporal data, the indexing system of Hadoop is not efficient for processing this type of data. Researchers attempted to design an indexing solution to make querying and processing convenient [3], but more research must be made into specialized DL architectures.

Time. Although data mining operations on climate data can take about three hours to process one terabyte of data before analyzing its spatiotemporal characteristics [3], one challenge with climate data is the need for large interval time series data spanning years before analysis [1]. Spatiotemporal

[8] http://climatedataguide.ucar.edu

analysis relies on observations over many years, and a period of 10 years will provide sufficient volumes of data but will not be comparable to the same interval in geological time.

Machine learning. Big Climate data has inherent differences with data suitable for more traditional data science: most climate data is organized in grids of spatiotemporal data where correlations happen autonomously when data is close in territory or proximity, causing independent variables to hold less practical information among data points; climate data responds to other types of evaluation metrics and data that is not typically analyzed by data science; and appropriate volumes of climate data has not been widely accessible to scientists and analysts until recently [1].

6 Further Research

For large volumes of open access environment and energy data to be widely available, there must be more research into the systems that can provide a proper canvas for analytics. Identified areas of research are in data formatting and control [5], i.e., in the building sector which has a large influence on energy consumption, more research effort is needed on how to collate different data sources and provide an architecture for their storage and facilitation. In terms of data quality issues of DL systems, other methodologies can be explored, such as lean supply chain management, that will further reduce expenses of DL [13].

Furthermore, hybrid systems that use DL concepts alongside other beneficial technologies are being studied: researchers in the industry are buzzing about a Data Lakehouse concept that will eventually take over both DL and DWH. Data lakehouses support the low-cost and heterogeneous storage of data lakes while allowing the on-demand analytical processing power of data warehouses [18]; similarly, the concept of ‘delta lakes’ replicate the low-cost high-volume storage of data lakes while allowing ACID transactions over tabular datasets and allowing time travel (undo schema rewrites) on query operations. [27]. Through proper research, these hybrid systems could manage to overcome the main challenges involved with implementing DL for climate and energy.

7 Conclusion

The main sentiment guiding this report is the need to provide unlimited access to those who can transform data into actionable knowledge that benefits societies. Due to their low-cost and high-output structure, DL are an innovative way to bridge the gap between public and private sector entities in the climate and energy sectors, where data scientists are currently faced with the obstacle of transforming information into knowledge. DL technology could be an answer to the question: how do we blend massive volumes of data from heterogenous sources while trading redundancy for complexity?

Through the perspective of sustainability, this paper sought to analyze the concept of DL as they are used in modern cases, and how DL can be supplemented with technology to increase their efficiency and innovate both Big Climate and renewable Big Energy, while open access to environment and energy data can provide a sufficient platform for DL to be supported. The challenges of each topic were presented in ways that further research will work to overcome.

Acronyms

AEP Annual Energy Production

DBMS Database Management System

DL Data Lake

DWH Data Warehouse

ETL/ELT Extract-Transfer-Load

GEMMS Generic and Extensible Metadata Management System

RDBMS Relational Database Management System

Tsdat Time Series Data Pipelines

UNFCC United Nations Framework Convention on Climate Change

References

[1] Faghmous, J. H., & Kumar, V. (2014). A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science. *Big Data*, 2(3), 155–163. https://doi.org/10.1089/big.2014.0026

[2] Suciu, G., Vulpe, A., Martian, A., Halunga, S., & Vizireanu, D. N. (2016). Big Data Processing for Renewable Energy Telemetry Using a Decentralized Cloud M2M System. *Wireless Personal Communications*, 87(3), 1113–1128. https://doi.org/10.1007/s11277-015-2527-7

[3] Li, Z., Hu, F., Schnase, J. L., Duffy, D. Q., Lee, T., Bowen, M. K., & Yang, C. (2017). A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce. *International Journal of Geographical Information Science*, 31(1), 17–35. https://doi.org/10.1080/13658816.2015.1131830

[4] Sarramia, D., Claude, A., Ogereau, F., Mezhoud, J., & Mailhot, G. (2022). CEBA: A Data Lake for Data Sharing and Environmental Monitoring. *Sensors*, 22(7), 2733. https://doi.org/10.3390/s22072733

[5] Mathew, P. A., Dunn, L. N., Sohn, M. D., Mercado, A., Custudio, C., & Walter, T. (2015). Big-data for building energy performance: Lessons from assembling a very large national database of building energy use. *Applied Energy*, 140, 85–93. https://doi.org/10.1016/j.apenergy.2014.11.042

[6] Fang, H. (2015). Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, 820–824. https://doi.org/10.1109/CYBER.2015.7288049

[7] Lansing, C., Levin, M., Sivaraman, C., Fao, R., & Driscoll, F. (2021). Tsdat: An Open-Source Data Standardization Framework for Marine Energy and Beyond. *OCEANS 2021: San Diego – Porto*, 1–6. https://doi.org/10.23919/OCEANS44145.2021.9706101

[8] Publications Office of the European Union., Capgemini Invent., Intrasoft International., Fraunhofer Fokus., con.terra., Sogeti., 52North., Time.lex., The Lisbon Council., & The University of Southampton. (2020). Reusing open data: A study on companies transforming open data into economic and societal value. Publications Office. https://data.europa.eu/doi/10.2830/876679

[9] Giest, S. (2017). Big data analytics for mitigating carbon emissions in smart cities: Opportunities and challenges. *European Planning Studies*, 25(6), 941–957. https://doi.org/10.1080/09654313.2017.1294149

[10] Perr-Sauer, J., Optis, M., Fields, J., Bodini, N., Lee, J., Todd, A., Simley, E., Hammond, R., Phillips, C., Lunacek, M., Kemper, T., Williams, L., Craig, A., Agarwal, N., Sheng, S., & Meissner, J. (2021). OpenOA: An Open-Source Codebase For Operational Analysis of Wind Farms. *Journal of Open Source Software*, 6(58), 2171. https://doi.org/10.21105/joss.02171

[11] Mack, P. (2014). Big Data, Data Mining, and Predictive Analytics and High Performance Computing. *Renewable Energy Integration*, pp. 439–454. https://doi.org/10.1016/B978-0-12-407910-6.00035-1

[12] Potdar, V., Chandan, A., Batool, S., & Patel, N. (2018). Big Energy Data Management for Smart Grids—Issues, Challenges and Recent Developments. In Z. Mahmood (Ed.), *Smart Cities* (pp. 177–205). Springer International Publishing. https://doi.org/10.1007/978-3-319-76669-0_8

[13] Derakhshannia, M., Gervet, C., Hajj-Hassan, H., Laurent, A., & Martin, A. (2020). Data Lake Governance: Towards a Systemic and Natural Ecosystem Analogy. *Future Internet*, 12(8), 126. ttps://doi.org/10.3390/fi12080126

[14] Hai, R., Quix, C., & Jarke, M. (2021). Data lake concept and systems: A survey. ArXiv:2106.09592 [Cs]. http://arxiv.org/abs/2106.09592

[15] Suleykin, A., Bobkova, A., Panfilov, P., & Chumakov, I. (2021). Efficient Data Exchange Between Typical Data Lake and DWH Corporate Systems. *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, 1–6. https://doi.org/10.1109/ICECET52533.2021.9698468

[16] Madera, C., & Laurent, A. (2016). The next information architecture evolution: The data lake wave. *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, 174–180. https://doi.org/10.1145/3012071.3012077

[17] Cuzzocrea, A. (2021). Big Data Lakes: Models, Frameworks, and Techniques. *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 1–4. https://doi.org/10.1109/BigComp51126.2021.00010

[18] Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. 8.

[19] Deligiannis, K., Raftopoulou, P., Tryfonopoulos, C., Platis, N., & Vassilakis, C. (2020). Hydria: An Online Data Lake for Multi-Faceted Analytics in the Cultural Heritage Domain. *Big Data and Cognitive Computing*, 4(2), 7. https://doi.org/10.3390/bdcc4020007

[20] Walker, C., & Alrehamy, H. (2015). Personal Data Lake with Data Gravity Pull. *2015 IEEE Fifth International Conference on Big Data and Cloud Computing*, 160–167. https://doi.org/10.1109/BDCloud.2015.62

[21] Ciampittiello, M., Manca, D., Dresti, C., Grisoni, S., Lami, A., & Saidi, H. (2021). Meteo-Hydrological Sensors within the Lake Maggiore Catchment: System Establishment, Functioning and Data Validation. *Sensors*, 21(24), 8300. https://doi.org/10.3390/s21248300

[22] Poudel, M., Sarode, R. P., Watanobe, Y., Mozgovoy, M., & Bhalla, S. (2022). Processing Analytical Queries over Polystore System for a Large Astronomy Data Repository. *Applied Sciences*, 12(5), 2663. https://doi.org/10.3390/app12052663

[23] European Commission. Directorate General for the Information Society and Media., Capgemini Consulting., Intrasoft International., Fraunhofer Fokus., con.terra., Sogeti., Open Data Institute., Time.lex., & University of Southampton. (2015). Creating value through open data: Study on the impact of re use of public data resources. Publications Office. https://data.europa.eu/doi/10.2759/328101

[24] Ceci, M., Corizzo, R., Fumarola, F., Ianni, M., Malerba, D., Maria, G., Masciari, E., Oliverio, M., & Rashkovska, A. (2014). Big Data Techniques For Supporting Accurate Predictions of Energy Production From Renewable Sources. Proceedings of the 19th International Database Engineering & Applications Symposium on - IDEAS ’15, 62–71. https://doi.org/10.1145/2790755.2790762

[25] Guo, H.-D., Zhang, L., & Zhu, L.-W. (2015). Earth observation big data for climate change research. *Advances in Climate Change Research*, 6(2), 108–117. https://doi.org/10.1016/j.accre.2015.09.007

[26] Schnase, J. L., Duffy, D. Q., Tamkin, G. S., Nadeau, D., Thompson, J. H., Grieg, C. M., McInerney, M. A., & Webster, W. P. (2017). MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service. *Computers, Environment and Urban Systems*, 61, 198–211. https://doi.org/10.1016/j.compenvurbsys.2013.12.003

[27] Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., Torres, J., van Hovell, H., Ionescu, A., Łuszczak, A., Świtakowski, M., Szafrański, M., Li, X., Ueshin, T., Mokhtar, M., Boncz, P., Ghodsi, A., Paranjpye, S., Senster, P., … Zaharia, M. (2020). Delta lake: High-performance ACID

table storage over cloud object stores. *Proceedings of the VLDB Endowment* 13(12), 3411–3424. https://doi.org/10.14778/3415478.3415560

[28] Huang, C.-C., Tseng, T.-L. B., & Zhou, M.-X. (2015). The novel rule induction approach to dynamic big data in green energy. *2015 Science and Information Conference (SAI)*, 1427–1432. https://doi.org/10.1109/SAI.2015.7237334