Metodologias e ferramentas de governança de dados aplicadas ao gerenciamento de data lakes: uma revisão sistemática

Wyllyany C. Santos[1], **David H. S. Lima**[1], **Carlos A. F. Silva**[1], **Márcio R. C. Ferro**[1]

[1]Instituto Federal de Alagoas - Campus Rio Largo – AL – Brasil

wcs11@aluno.ifal.edu.br

{david.lima, carlos.feitosa, marcio.roberio}@ifal.edu.br

Abstract. *A data lake is a centralized repository designed to store diverse types of data, regardless of format or structure. While this flexibility offers significant advantages, it also poses the risk of turning the repository into a data swamp that is a scenario where disorganized, inconsistent, or low-value data accumulates. To mitigate this risk, the adoption of robust data governance practices is essential to ensure proper organization and efficient data management. Given the existing knowledge gaps concerning effective governance implementation in data lakes, this study presents a systematic literature review aimed at identifying the methodologies and tools currently employed in managing such repositories.*

Resumo. *Um data lake é um repositório centralizado de grande porte, utilizado para armazenar dados de qualquer tipo, sem restrições quanto ao formato ou à estrutura. No entanto, essa flexibilidade pode resultar na formação de um pântano de dados que é uma situação em que o repositório passa a concentrar informações desorganizadas, inconsistentes ou de baixo valor. Para evitar esse cenário, torna-se essencial a adoção de práticas eficazes de governança de dados, que garantam o armazenamento adequado e a gestão eficiente das informações. Considerando as lacunas existentes na literatura sobre a implementação da governança em data lakes, este estudo propõe uma revisão sistemática da literatura com o objetivo de identificar metodologias e ferramentas utilizadas na gestão desses repositórios.*

1. Introdução

Governança de dados refere-se a um conjunto de práticas, políticas, processos e estruturas organizacionais voltados para assegurar o gerenciamento eficaz dos dados dentro de uma organização. Seu principal objetivo é garantir que os dados sejam precisos, acessíveis, seguros e utilizados de forma ética, em conformidade com normas e regulamentações vigentes [DAMA International 2017]. Segundo Derakhshannia et al. [Derakhshannia et al. 2020], a governança de dados está intrinsecamente ligada ao ciclo de vida dos dados, bem como à sua qualidade e segurança, independentemente do sistema de armazenamento utilizado. Para promover uma gestão eficiente, recorre-se ao Data Management Body of Knowledge (DAMA-DMBOK), um guia elaborado pela DAMA International, organização dedicada à disseminação de boas práticas em gerenciamento de dados e ao apoio a profissionais e instituições nessa área [Bližnák et al. 2024].

O DAMA-DMBOK é estruturado em torno de onze áreas de conhecimento, organizadas em um framework conhecido como Roda DAMA (Figura 1) [DAMA International 2017]. Neste estudo, essa estrutura será utilizada como

referência para identificar e categorizar as metodologias de governança de dados aplicadas à gestão eficiente de data lakes. A Roda DAMA posiciona a governança de dados no centro das atividades de gerenciamento, refletindo seu papel fundamental na promoção da consistência e no equilíbrio entre as diferentes funções. As demais áreas de conhecimento, como Arquitetura de Dados, Modelagem e Design de Dados, entre outras, são dispostas ao redor da governança, compondo uma abordagem integrada. Embora todas as áreas sejam essenciais para uma gestão de dados madura, sua implementação pode ocorrer de forma gradual, de acordo com as necessidades e prioridades de cada organização.

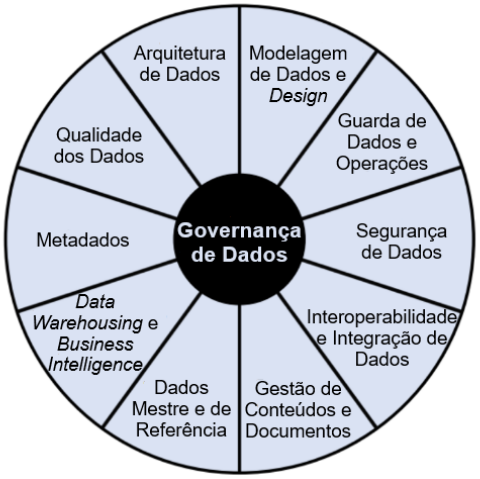

Figura 1. Áreas abordadas pela Roda Dama

Um data lake é um repositório centralizado de dados que, na ausência de mecanismos adequados de governança, pode rapidamente se transformar em um pântano de dados, caracterizado pelo acúmulo de informações desorganizadas e de baixo valor [Nambiar and Mundra 2022]. Diante desse cenário, este estudo realiza uma revisão sistemática da literatura com o objetivo de identificar metodologias e ferramentas de governança aplicadas à gestão de data lakes. A estrutura do artigo está organizada da seguinte forma: a Seção 2 apresenta a fundamentação teórica; a Seção 3 descreve a metodologia adotada; a Seção 4 expõe os resultados e a discussão; e, por fim, a Seção 5 traz as considerações finais.

2. Fundamentação Teórica

Big Data refere-se a grandes volumes de dados dinâmicos e heterogêneos, gerados por pessoas, dispositivos e sistemas automatizados. Esses dados são comumente caracterizados pelos cinco V's: Volume, Velocidade, Variedade, Veracidade e Valor [Ishwarappa and Anuradha 2015]. Dentro dessa infraestrutura, o data lake surge como um componente essencial, atuando como repositório para a ingestão, armazenamento e disponibilização de dados em larga escala. Sem o suporte de data lakes ou arquiteturas tecnológicas equivalentes, torna-se inviável enfrentar os desafios técnicos impostos pelas características do Big Data. Segundo a DAMA International [DAMA International 2017], soluções baseadas em Big Data, como os data lakes, fazem uso do modelo ELT (Extração, Carregamento e Transformação), que é mais adequado ao tratamento de grandes volumes de dados em comparação ao tradicional modelo ETL. A alta velocidade e o grande volume de dados exigem abordagens específicas para questões críticas do gerenciamento de

dados, como a integração, o gerenciamento de metadados e a avaliação da qualidade da informação.

De acordo com a DAMA International [DAMA International 2017], um data lake é um ambiente projetado para ingerir, armazenar, avaliar e analisar grandes volumes de dados com diferentes tipos e estruturas. Contudo, a ausência de práticas adequadas de gestão pode comprometer a organização, a integridade e a utilidade dos dados armazenados. Nesse contexto, o gerenciamento de metadados torna-se essencial para manter o controle sobre o conteúdo do repositório, especialmente durante a fase de ingestão.

A implementação de um data lake geralmente envolve uma infraestrutura composta por diversas ferramentas e serviços especializados em armazenamento e processamento de dados, como o Hadoop, além de mecanismos de transformação e integração. Entre as plataformas utilizadas destacam-se Amazon S3 (AWS) [1], Azure Data Lake Storage (ADLS) [2], Google Cloud Storage (GCS) [3], Hadoop Distributed File System (HDFS) [4] e Databricks Lakehouse [5].

3. Metodologia

Segundo Galvão e Ricarte [Galvão and Ricarte 2020], o conceito de revisão de literatura é amplo, abrangendo diversos tipos de estudos que analisam produções acadêmicas sobre temas específicos. Este estudo adota a revisão sistemática como abordagem metodológica, com base nas diretrizes apresentadas por Galvão e Ricarte [Galvão and Ricarte 2020]. Trata-se de um tipo de pesquisa que segue protocolos rigorosos e visa organizar, de forma lógica e estruturada, um extenso corpus documental. As etapas envolvidas incluem: definição da questão de pesquisa, seleção das bases de dados, elaboração da estratégia de busca, triagem e seleção dos documentos, e sistematização dos resultados.

3.1. Delimitação da questão de pesquisa

A definição da questão de pesquisa é fundamental para orientar todas as etapas do estudo, pois consiste em uma pergunta clara, delimitada e passível de ser respondida por meio de métodos científicos. Neste trabalho, a questão proposta é: "Quais são as metodologias e ferramentas de governança de dados aplicadas no gerenciamento de data lakes?". Para sua formulação, foi utilizado o modelo PICO, estruturado da seguinte forma: P (População ou problema) – gerenciamento de data lakes; I (Intervenção) – governança de dados; C (Comparação) – não se aplica neste contexto; e O (Outcome, ou desfecho esperado) – metodologias e ferramentas utilizadas.

3.2. Seleção das bases de dados

Com a questão de pesquisa devidamente definida, o próximo passo consiste na seleção das bases de dados nas quais será realizada a busca por artigos relevantes. Embora haja

[1] https://aws.amazon.com/
[2] https://azure.microsoft.com/
[3] https://cloud.google.com/
[4] https://hadoop.apache.org/
[5] https://www.databricks.com/

possibilidade de sobreposição de resultados entre diferentes bases, cada uma possui características específicas, direcionadas a públicos distintos e com maior ênfase em determinadas áreas do conhecimento. Nesta revisão sistemática, foram utilizadas as seguintes bases: ACM Digital Library [6], Elsevier [7], Google Scholar [8] e IEEE Xplore [9].

3.3. Elaboração da estratégia de busca

Neste estudo, a busca foi realizada em 3 de novembro de 2024, contemplando publicações no período de 2020 a 2024. Foi escolhido esse período para contemplar artigos recentes, relacionados à essa revisão. Utilizou-se a seguinte string de busca: ("Data Lake Governance") AND (Methodology OR Methodologies OR Tools OR Tool). A consulta resultou em um total de 243 artigos, distribuídos da seguinte forma: 93 provenientes do Google Scholar, 10 do IEEE Xplore, 128 da ACM Digital Library e 12 da Elsevier.

3.4. Seleção dos documentos

Numa revisão sistemática, o processo de seleção dos estudos deve ser conduzido por, no mínimo, dois pesquisadores, que aplicam os critérios de inclusão e exclusão previamente definidos. Os critérios de inclusão determinam os requisitos que um artigo deve obrigatoriamente atender para ser considerado na análise, enquanto os critérios de exclusão especificam as condições que justificam a eliminação de um estudo. Em casos de discordância, a decisão final deve ser tomada por um terceiro pesquisador.

Nesta revisão, foram adotados os seguintes critérios de inclusão: IC1 – Artigos que abordam a governança em data lakes e IC2 – Artigos que apresentam metodologias ou ferramentas para implementar a governança em data lakes. Os critérios de exclusão aplicados foram: EC1 – Literatura cinzenta, incluindo documentos não revisados por pares, como relatórios técnicos, livros, teses ou dissertações; EC2 – Artigos secundários que apenas realizam revisões temáticas, como revisões sistemáticas da literatura ou surveys; EC3 – Artigos não escritos em inglês; EC4 – Artigos de extensão reduzida, como pôsteres, resumos expandidos ou banners; EC5 – Artigos redundantes de um mesmo autor, mantendo-se apenas o mais completo e recente; EC6 – Estudos que não atendem a pelo menos um dos critérios de inclusão estabelecidos.

Para a triagem das referências, foi utilizada a plataforma Rayyan, em sua versão gratuita, e algumas foram importadas no formato RIS e outras no BibTeX. O Rayyan realizou a detecção automática de registros duplicados, possibilitando a revisão manual e a exclusão desses itens. A seleção dos documentos foi conduzida com base nos critérios de inclusão e exclusão previamente definidos, levando em consideração o tipo de documento, o título e o resumo.

Ao final do processo, 28 artigos foram incluídos na análise, todos atendendo ao critério de inclusão 1 (IC1), dos quais 15 também satisfaziam o critério de inclusão 2 (IC2). No entanto, apenas 12 artigos puderam ser lidos integralmente, em razão de restrições de acesso aos demais. Os documentos excluídos foram classificados da seguinte forma: 8 duplicados, 23 por EC1, 36 por EC2, 6 por EC3, 2 por EC5 e 140 por EC6. Nenhum artigo foi excluído com base no critério EC4.

[6] https://dl.acm.org/
[7] https://www.elsevier.com/
[8] https://scholar.google.com.br/
[9] https://ieeexplore.ieee.org/

3.5. Sistematização dos resultados

Após a seleção dos textos, a equipe responsável pela revisão deve proceder à leitura integral dos estudos incluídos, com o objetivo de extrair informações comparáveis entre si. Entre os dados a serem coletados, destacam-se: data e país de realização do estudo, população ou contexto analisado, intervenção proposta, método empregado e principais resultados ou desfechos observados. Essas informações podem ser organizadas em quadros ou tabelas, de modo a facilitar a análise e a apresentação dos achados no relatório final da revisão.

4. Resultado e discussão

Dos 12 artigos analisados, todos abordaram pelo menos uma das áreas de conhecimento da Roda DAMA, sendo que alguns estudos contemplaram até cinco áreas simultaneamente — como é o caso da construção de data lake apresentada por Sarramia et al. [Sarramia et al. 2022]. Por outro lado, alguns artigos concentraram-se em uma única área ou ferramenta específica, como observado em Gyulgyulyan e Astsatryan [Gyulgyulyan and Astsatryan 2023].

A Tabela 1 apresenta a relação entre os artigos selecionados e as respectivas áreas da Roda DAMA exploradas em cada estudo. Para facilitar a leitura, as áreas foram codificadas conforme a legenda a seguir: 1 – Governança de Dados (Data Governance), 2 – Arquitetura de Dados (Data Architecture), 3 – Modelagem e Design de Dados (Data Modeling and Design), 4 – Armazenamento e Operações de Dados (Data Storage and Operations), 5 – Segurança de Dados (Data Security), 6 – Integração e Interoperabilidade de Dados (Data Integration and Interoperability), 7 – Gerenciamento de Documentos e Conteúdo (Document and Content Management), 8 – Dados Mestres e de Referência (Reference and Master Data), 9 – Data Warehousing e Business Intelligence, 10 – Metadados (Metadata), 11 – Qualidade de Dados (Data Quality).

O estudo de Giebler et al. [Giebler et al. 2020] propõe um modelo de gerenciamento de data lakes baseado em zonas de processamento (brutas, limpas e agregadas), com regras de governança relacionadas a acesso, responsabilidades e qualidade dos dados, associadas às áreas 1 e 11 da Roda DAMA. Para lidar com a falta de padronização, os autores desenvolveram um meta-modelo de zonas, validado por protótipo em cenário real, demonstrando sua aplicabilidade e eficiência na gestão de data lakes.

O artigo de Hamadou et al. [Hamadou et al. 2020] aborda o Data Lake Nacional de Energia da Dinamarca (FEDDL), concebido como uma base para a coleta, compartilhamento e análise de grandes volumes de dados relacionados ao setor energético. A infraestrutura do FEDDL tem como objetivo otimizar o uso flexível de fontes de energia renováveis, contando com o suporte de tecnologias como inteligência artificial e aprendizado de máquina. O estudo descreve os requisitos e a arquitetura do sistema, estruturada em múltiplas camadas incluindo ingestão, armazenamento, exploração, governança e privacidade. A seleção das ferramentas utilizadas foi criteriosamente justificada, com destaque para o uso do Apache Atlas como mecanismo de governança de metadados. Um dos principais desafios identificados no projeto foi a proteção de dados sensíveis. As contribuições do estudo estão relacionadas, principalmente, às áreas 2 e 10 da Roda DAMA.

Tabela 1. Artigos selecionados e lidos

Título	Ano	Ferramentas	METODOLOGIAS									
			1	2	3	4	5	6	7	8	9	10	11
A Zone Reference Model for Enterprise-Grade Data Lake Management	2020	Gerenciamento de data lake baseado em zonas	✓										✓
The Danish National Energy Data Lake	2020	Apache Atlas para metadados		✓								✓	
Data Lake: A Case of Study of a Big Data Analytics Architecture for Public Procurements	2021	Gerenciamento de metadados usando Delta Lake							✓			✓	
DataOps for Cyber-Physical Systems Governance	2021	DataOps				✓							
Huawei and International Data Spaces	2022	GAIA-X	✓	✓			✓	✓					
CEBA: A Data Lake for Data Sharing and Environmental Monitoring	2022	FAIR	✓					✓		✓	✓	✓	
Alert System for Data Quality in Data Lakes	2023	Alertas											✓
TEADAL: Trustworthy, Energy-Aware federated Data Lakes	2023	TEADAL				✓	✓	✓					
Big Data System for Regional Lakes in CBM Development	2023	Teoria do lago de dados regional											✓
Design Resilient Data Lakes with Biology and CS	2023	Redes ecológicas e Teoria dos Grafos	✓										
IBM-Watson knowledge catalog	2023	IBM-WKC		✓	✓				✓				
Enhancing Data Lake Management with LDA Approach	2024	Algoritmo LDA							✓				

O artigo de Sosa e Paciello [Sosa and Paciello 2021] analisa os desafios do Big Data no tratamento de grandes volumes de dados, utilizando dados abertos de compras públicas como estudo de caso. Os autores comparam a ferramenta relacional KingFisher com o Delta Lake, baseado em Apache Spark, e demonstram que este último reduz significativamente o uso de armazenamento, sendo mais eficiente em cenários com alto volume de dados. Concluem que o data lake é uma alternativa promissora para ambientes de Big Data. Embora não trate diretamente da governança de dados, o estudo aborda o gerenciamento escalável de metadados, relacionado às áreas 7 e 10 da Roda DAMA.

O artigo de Garriga et al [Garriga et al. 2021] adota o conceito de DataOps, um conjunto de práticas e processos voltados à operacionalização do ciclo de vida dos dados. O foco do DataOps está na automação, integração e entrega contínua de dados confiáveis, com o objetivo de apoiar análises e tomadas de decisão em tempo hábil. O estudo propõe uma metodologia de análise preditiva baseada em Sistemas Ciberfísicos, aliada a um pipeline de dados orientado por princípios de DataOps, visando aprimorar as operações de companhias aéreas e aeroportos. Essa abordagem está diretamente relacionada à área 4

da Roda DAMA.

O estudo de O'Brien et al. [O'Brien et al. 2022] destaca a importância da Integração e Interoperabilidade de Dados em data lakes, especialmente diante dos desafios enfrentados por empresas europeias na digitalização de processos. Iniciativas como o International Data Spaces (IDS) e o GAIA-X promovem soberania e segurança de dados, fundamentais para um ecossistema interoperável. Setores como saúde, manufatura e telecomunicações ainda sofrem com a fragmentação de dados, o que motivou a criação de políticas como a Estratégia Europeia para Dados. A Huawei contribui com arquiteturas abertas, alinhadas aos princípios do IDS, para facilitar o compartilhamento seguro de dados. A abordagem do estudo se relaciona às áreas 2, 5 e 6 da Roda DAMA.

O artigo de Sarramia et al. [Sarramia et al. 2022] apresenta o CEBA, uma plataforma institucional voltada ao armazenamento, compartilhamento e valorização de dados ambientais heterogêneos. Sua arquitetura permite o gerenciamento de diferentes tipos de dados com informações mínimas (o quê, onde e quando), integrando coordenadas geográficas em todo o ciclo de vida. Os dados são indexados e disponibilizados quase em tempo real, com aplicação em áreas como agricultura de precisão, biodiversidade e monitoramento ambiental. O CEBA adota os princípios FAIR e a iniciativa INSPIRE, armazenando os dados em um data lake acessado via catálogo espacial. A governança é baseada em boas práticas distribuídas entre os componentes da arquitetura, com destaque para o gerenciamento de metadados. O estudo contempla as áreas da Roda DAMA: 1, 6, 8, 9 e 10.

O trabalho de Gyulgyulyan e Astsatryan [Gyulgyulyan and Astsatryan 2023] propõe um sistema de alerta voltado à qualidade dos dados, capaz de identificar falhas durante análises em tempo real. A solução contribui para decisões mais precisas e para a redução de custos e riscos operacionais. Já o estudo de Plebani et al. [Plebani et al. 2023] descreve o projeto TEADAL, que busca implementar federações de data lakes confiáveis, energeticamente eficientes e alinhadas com princípios de privacidade e soberania de dados. A proposta visa facilitar o compartilhamento de dados entre organizações ao longo do continuum computacional (borda, névoa e nuvem), abrangendo as áreas 5 e 6 da Roda DAMA.

O estudo de Wang et al. [Wang et al. 2023] apresenta o desenvolvimento e a aplicação de um modelo de governança de dados voltado à gestão de informações sobre o metano de carvão (CBM), fundamentado na teoria de data lake regional. A proposta inclui a definição de um modelo de otimização para diferentes tipos de dados, a construção de um modelo de expansão com base nas características geológicas dos reservatórios e o estabelecimento de um acoplamento entre dados de campo, laboratoriais, gerenciais e o sistema de dados. Um dos principais resultados do estudo é a estrutura de governança proposta, composta por quatro componentes: suporte básico, ciclo de vida dos dados, áreas centrais e suporte estratégico. A área da Roda DAMA contemplada é a 11.

O artigo de Derakhshannia et al. [Derakhshannia et al. 2023] promove uma abordagem interdisciplinar ao aplicar conceitos da biologia na ciência da computação e ciência de dados, com o objetivo de propor estratégias de resiliência para data lakes. Embora a contribuição direta ao campo biológico tenha sido limitada, o estudo gerou importantes reflexões sobre a necessidade de avaliar a eficácia das estratégias propostas na arquitetura

técnica desses repositórios. O foco principal da pesquisa recai sobre a área 2 da Roda DAMA.

O estudo de Cherradi et al. [Cherradi et al. 2023] apresenta a análise de um catálogo de dados robusto, desenvolvido para aprimorar a gestão de fontes heterogêneas em data lakes. O trabalho destaca a aplicação de técnicas de processamento de linguagem natural no IBM Watson Knowledge Catalog, com foco nas áreas 3 e 7 da Roda DAMA.

O artigo de Cherradi e Haddadi [Cherradi and El Haddadi 2024] demonstra a aplicação da técnica Latent Dirichlet Allocation (LDA) como estratégia para organizar e compreender melhor os dados armazenados em data lakes. O LDA é um método de modelagem temática que agrupa palavras e documentos com base em tópicos, facilitando a recuperação e o uso eficiente das informações. Os resultados do estudo indicam que a técnica contribui significativamente para evitar a degradação do repositório em um pântano de dados, promovendo maior agilidade na localização de conteúdos relevantes. A área da Roda DAMA contemplada é 7.

5. Considerações finais

A revisão sistemática realizada evidenciou uma lacuna na literatura quanto à aplicação integrada de todas as áreas da Roda DAMA no contexto do gerenciamento de data lakes. Embora nem todo ambiente exija a implementação plena de todas essas áreas, sua integração é considerada essencial para alcançar um modelo de governança de dados maduro e eficaz. Entre os temas mais recorrentes nos estudos analisados destacam-se: Arquitetura de Dados, Integração e Interoperabilidade, Gerenciamento de Documentos e Conteúdos, Metadados e Qualidade de Dados, conforme apresentado na Tabela 1.

Esses resultados estão alinhados com a literatura especializada, que reconhece a integração de dados, o gerenciamento de metadados e a garantia da qualidade da informação como pilares centrais na administração de data lakes. Como diretriz para investigações futuras, recomenda-se o desenvolvimento de soluções que incorporem, de forma articulada, todas as áreas da Roda DAMA, mesmo reconhecendo os desafios associados à adaptação dessas práticas a diferentes contextos organizacionais e necessidades específicas.

Referências

Bližnák, K., Munk, M., and Pilková, A. (2024). A systematic review of recent literature on data governance (2017–2023). *IEEE Access*, 12:149875–149888.

Cherradi, M., Bouhafer, F., and Haddadi, A. E. (2023). Data lake governance using ibm-watson knowledge catalog. *Scientific African*, 21.

Cherradi, M. and El Haddadi, A. (2024). Enhancing data lake management systems with lda approach. *Journal of Data Science and Intelligent Systems*, 3(1):58–66.

DAMA International (2017). *DAMA-DMBOK: Data Management Body of Knowledge*. Technics Publications, USA, 2 edition.

Derakhshannia, M., Gervet, C., Hajj-Hassan, H., Laurent, A., and Martin, A. (2020). Data lake governance: Towards a systemic and natural ecosystem analogy. *Future Internet*, 12:1–16.

Derakhshannia, M., Laurent, A., and Martin, A. (2023). Mixing biology and computer science concepts to design resilient data lakes. *Journal of Interdisciplinary Methodologies and Issues in Science*, 11.

Galvão, M. C. B. and Ricarte, I. L. M. (2020). Revisão sistemática da literatura: conceituação, produção e publicação. *LOGEION: Filosofia da informação*, 6:57–63.

Garriga, M., Aarns, K., Tsigkanos, C., Tamburri, D. A., and Heuvel, W. V. D. (2021). Dataops for cyber-physical systems governance: The airport passenger flow case. *ACM Transactions on Internet Technology*, 21(2):Article 36, 25 pages.

Giebler, C., Gröger, C., Hoos, E., Schwarz, H., and Mitschang, B. (2020). A zone reference model for enterprise-grade data lake management. In *2020 IEEE 24th International Enterprise Distributed Object Computing Conference (EDOC)*, pages 57–66, Eindhoven, Netherlands.

Gyulgyulyan, E. and Astsatryan, H. (2023). Alert system for data quality in data lakes. In *CSIT Conference 2023*, Yerevan, Armenia.

Hamadou, H. B., Bach Pedersen, T., and Thomsen, C. (2020). The danish national energy data lake: Requirements, technical architecture, and tool selection. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1523–1532, Atlanta, GA, USA.

Ishwarappa and Anuradha, J. (2015). A brief introduction on big data 5vs characteristics and hadoop technology. *Procedia Computer Science*, 48:319–324.

Nambiar, A. and Mundra, D. (2022). An overview of data warehouse and data lake in modern enterprise data management. *Big Data and Cognitive Computing*, 6(4):132.

O'Brien, M. A., Mohally, D., Brasche, G. P., and Sanfilippo, A. G. (2022). Huawei and international data spaces. In Otto, B., ten Hompel, M., and Wrobel, S., editors, *Designing Data Spaces*. Springer, Cham.

Plebani, P., Kat, R., Pallas, F., Werner, S., Inches, G., Laud, P., and Santiago, R. (2023). Teadal: Trustworthy, energy-aware federated data lakes along the computing continuum. In *CEUR Workshop Proceedings*, volume 3413, pages 28–35.

Sarramia, D., Claude, A., Ogereau, F., Mezhoud, J., and Mailhot, G. (2022). Ceba: A data lake for data sharing and environmental monitoring. *Sensors*, 22:2733.

Sosa, D. and Paciello, J. (2021). Data lake: A case of study of a big data analytics architecture for public procurements. In *2021 Eighth International Conference on eDemocracy & eGovernment (ICEDEG)*, pages 194–198, Quito, Ecuador.

Wang, H., Adenutsi, C. D., Wang, C., Sun, Z., Zhang, Y., Li, Y., Zhang, Z., and Wang, J. (2023). Construction and application of a big data system for regional lakes in coalbed methane development. *ACS Omega*, 8(20):18323–18331.