Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

An Integrated Hybrid Lakehouse Framework for Data Governance, Data Lineage, and Enterprise Analytics

Lokeshkumar Madabathula

Independent Researcher, San Antonio, Texas, USA

ARTICLE INFO

Received: 04 Nov 2024

Revised: 17 Dec 2024

Accepted: 28 Dec 2024

ABSTRACT

Present-day companies gather data from various databases, cloud computing, applications, and analytics. Nonetheless, the problem arises when a company has fragmented data which leads to problems such as data governance, data lineage, data quality, accessibility, and analytical trust. This research offers a conceptual model of a lakehouse solution with integrated data governance, data lineage, and analytics for the enterprise. The proposed methodology is qualitative, inductive, and based on secondary data. The secondary sources consist of articles from peer-reviewed journals, conference papers, technical reports, industry reports, and organizational reports. The literature review consists of concepts associated with Lakehouse Architecture, Data Governance, Metadata Management, Data Lineage, Data Quality, Cloud Computing, Business Intelligence, Real-Time Analytics, Artificial Intelligence, and Machine Learning. The findings prove that lakehouse architecture supports scalable storage, flexible processing, open format interoperability, and various analytical workloads. Still, there are several technical and operational challenges. These gaps have been addressed in the suggested framework by means of centralized metadata, automated governance, lineage from end to end, standardized data, quality assurance, and constant monitoring. The combination of these features may contribute to higher transparency, reliability, availability, and consistency of data analysis. Moreover, business intelligence, real-time analytics, AI, and machine learning can be enabled by the usage of data which has been governed and traceable. Thus, the research shows how the combination of governance and lineage with the help of lakehouse design pattern can improve enterprise analytics. This solution eliminates unnecessary duplications in data platforms while improving the visibility throughout the whole data lifecycle.

Keywords: Lakehouse Architecture; Data Governance; Data Lineage; Enterprise Analytics; Metadata; Data Quality; Cloud Computing; Business Intelligence; Data Integration; Decision Support

INTRODUCTION

Despite huge amounts of data collected by modern corporations from databases, applications, clouds, and connected devices, fragmented data landscapes can give rise to such problems as data quality, ownership, and data trust. Lakehouse architecture involves combining the flexibility of data lakes with the management capabilities of data warehouses. In this regard, the lakehouse architecture provides scalable storage while allowing for effective analytics using structured and unstructured data. Data governance involves defining issues regarding ownership, access to the data, data quality, security of the data, and appropriate use of data within enterprise environments (Janssen et al., 2020). Data lineage entails identifying the source of the data, modifications made to the data, and consumption of the data. These functionalities enhance transparency and accountability as far as enterprise analytics is concerned. Integrating governance and lineage into such environments is a difficult task. The current research proposes an integrated hybrid lakehouse architecture that incorporates governance, lineage, and enterprise analytics.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

METHOD

The current research will employ a qualitative approach since it explores the concepts, practices and relationships within the architecture of lakehouse. The use of a qualitative approach for the current study is appropriate since the research is aimed at exploring governance and lineage for the purpose of enterprise analytics (Kheyrossadat, 2020). The current study employs an inductive approach, whereby the study develops generalizations based on the available evidence. This approach helps the study to explore various practices and then make generalizations on the integrated lakehouse approach. The current research employs secondary data collection since there is no need for any primary data collection in the form of survey or interview (Taherdoost, 2021). Data is collected from peer-reviewed journal articles, conference papers, industry reports, technical documents and other reliable organizational publications. The data sources will include lakehouse architecture, data governance, data lineage, metadata management, data integration, data quality, cloud data platforms and enterprise analytics. These sources offer proven knowledge and evidence for assessing the existing approaches (Bennouna et al., 2021). The results are then analyzed thematically for any common problems, strengths, and needs for integration. This helps develop a practical hybrid framework from the existing research and industry practice.

RESULTS

Current Lakehouse Architecture and Integration Capabilities

From the analyzed literature, it is evident that lakehouse architecture incorporates the advantages of data lakes together with those of data warehouses. This is because it employs affordable cloud object storage for managing structured, semi-structured and unstructured data. The open data formats like Parquet, Delta Lake, Apache Iceberg, and Hudi enhance interoperability of data platforms. Besides, the architecture offers ACID transactions, schema management, data versioning, and time travel services (Armbrust et al., 2020). Such advantages eliminate major challenges associated with the classical data lakes including poor transactional management and data structuring.

Table 1: Technical Capabilities of the Proposed Hybrid Lakehouse

Technical Capability	Function	Data Type	Market Data	Key Benefit	Application
Lakehouse	Unified Storage	Mixed Data	67% Adoption	Lower Duplication	Enterprise Analytics
Object Storage	Data Storage	Raw Data	54.7% Cloud	Scalable Storage	Data Management
Delta Lake	ACID Management	Structured Data	Open Format	Data Reliability	Transaction Processing
Apache Iceberg	Table Management	Structured Data	Open Format	Interoperability	Data Analytics
Metadata Catalog	Data Discovery	Metadata	60% Governance Need	Better Visibility	Data Governance
Data Lineage	Traceability	Processed Data	48% Framework Need	Error Tracking	Impact Analysis
Batch Processing	Data Processing	Historical Data	Large Scale	Efficient Processing	BI Reporting
Stream Processing	Real-Time Analysis	Streaming Data	Real-Time	Faster Decisions	Live Monitoring
Business Intelligence	Data Visualisation	Structured Data	BI Growth	Better Insights	Dashboards
AI/ML Analytics	Predictive Analysis	Mixed Data	AI Growth	Forecasting	Decision Support

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ Research Article

Moreover, the analyzed evidence points out the capability of lakehouses to minimise data duplication due to the possibility of running multiple analytical workloads using common storage. Batch and real-time stream processing can be used in the architecture to support not only the reporting but also advanced analytics. Moreover, metadata catalogs have features to discover datasets, perform access control, classification and governance. The support for multiple engines allows SQL, Python, Spark, and other processing engines to use the same common datasets (Ordonez et al., 2020).

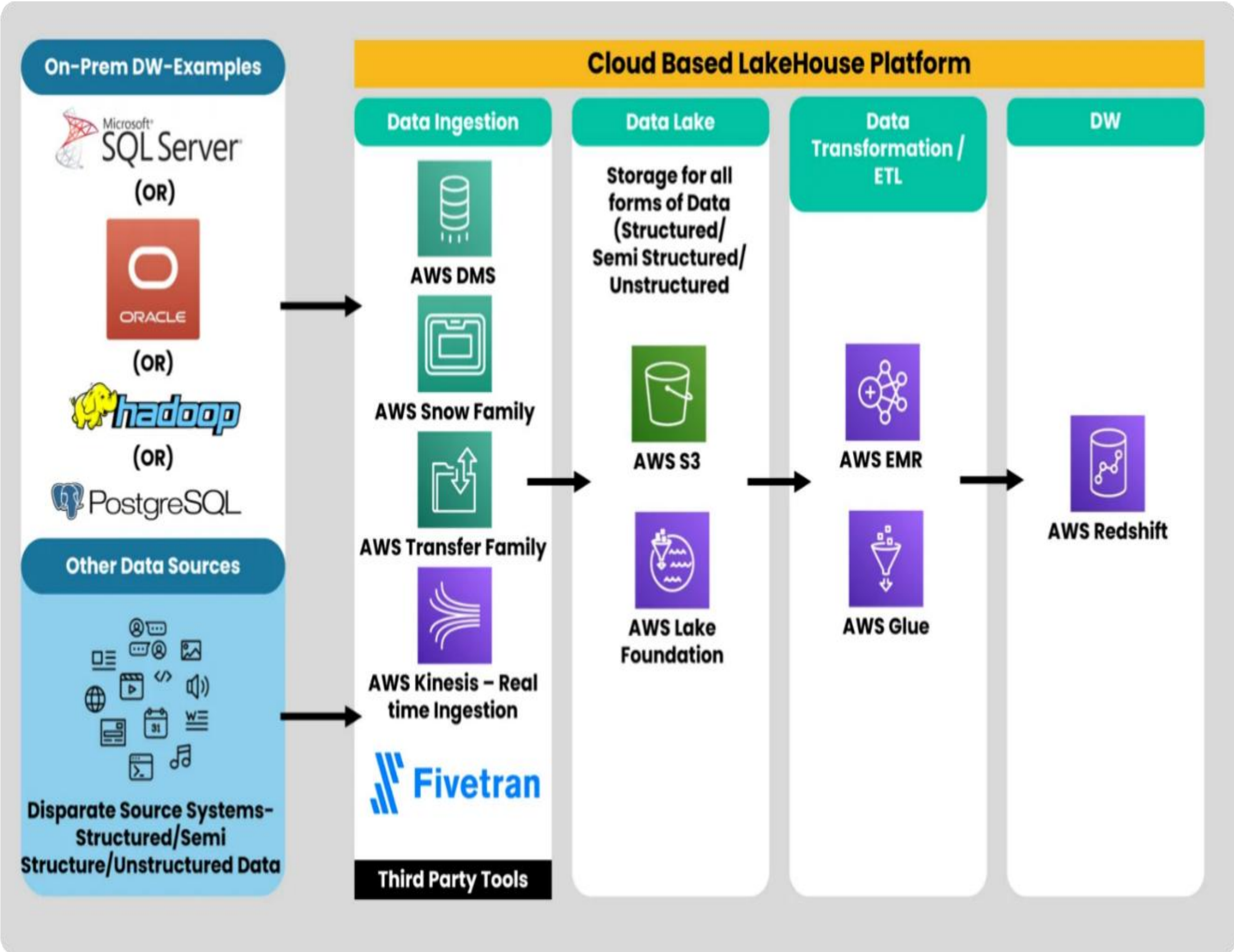

Figure 1: Lakehouse implemented on different cloud platforms

(Source: Subramanian, 2023)

In general, the results show that lakehouse architecture is a more consolidated basis for analytics compared to separate lake and warehouse architectures. However, based on the collected data, it is clear that even the integration is not going to guarantee efficient governance and data transparency. Hence, it would be fair to say that the above-mentioned aspects need to be integrated in the lakehouse model.

Data Governance and Quality Management Practices

The need for data governance is crucial in guaranteeing that enterprise data is accurate and protected from misuse and abuse. Data governance includes specifying who is the owner of the data, along with the processes, policies, and access rights of the data in the data ecosystem. Data quality management consists of managing the accuracy, completeness, consistency, timeliness, and validity of data. In the lakehouse architecture, data governance can be

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

accomplished using the controls of central data catalogues, data access, metadata management, and data quality management (Han, 2021).

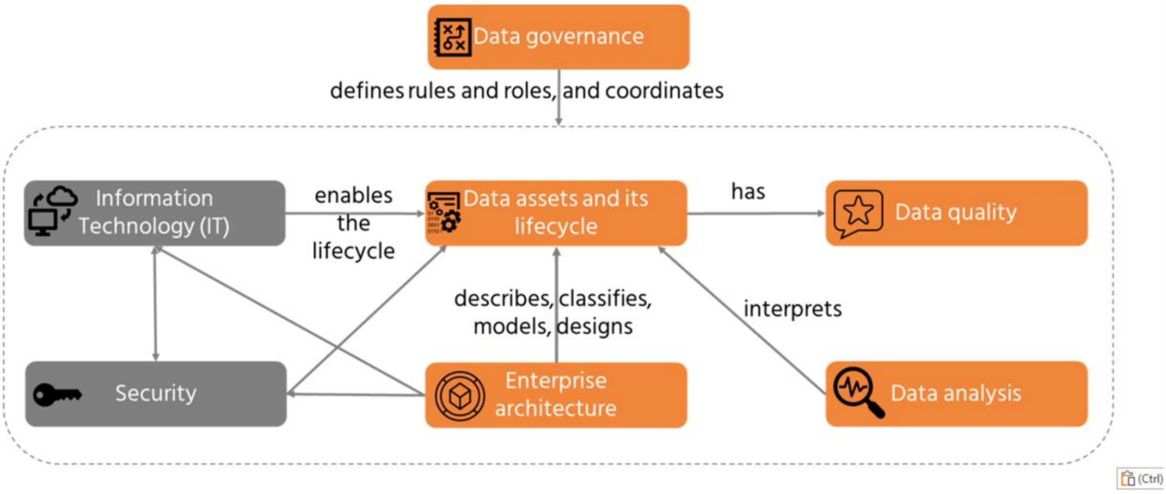

Figure 2: Concept map of key data management capabilities.

(Source: Steenbeek, 2023)

These controls help organizations in detecting problematic data before it reaches the analytic tools. Automated validation rules can even be used for monitoring the datasets in order to detect missing, duplicate, inconsistent, or outdated data. Governance processes can also help organizations comply with regulations through controlling access to sensitive data as well as maintaining records about data management activities. Nevertheless, the existence of governance in many different platforms might lead to inconsistencies in policies and duplication of controls (Staub et al., 2022). Thus, the evidence suggests incorporating governance and quality management within the lakehouse design.

Data Lineage and Metadata Visibility

Evidence provided in the review highlights that data lineage and metadata visibility are crucial to increasing trust in enterprise data. Data lineage captures the history of data movement from the source to processing, transformations, storage, and finally consumption. This enables enterprises to understand the process of generating analytics as well as the source of any potential error. Metadata refers to data about datasets such as the dataset schema, owner, location, quality, and use cases (Visengeriyeva & Abedjan, 2020). In a lakehouse environment, the metadata catalogues can offer a consolidated view of the datasets from different sources and processing platforms.

Table 2: Metadata and Data Traceability Capabilities

Data Source	Metadata Visibility	Lineage Tracking	Key Benefit	Governance Role	Analytics Impact
Databases	High	End-to-End	Better Traceability	Access Control	Trusted Results
Cloud Storage	Centralised	Automated	Faster Discovery	Data Classification	Better Decisions
Applications	Unified	Source-to-Target	Error Identification	Ownership	Consistent Reporting
Streaming Data	Real-Time	Continuous	Faster Monitoring	Quality Control	Timely Analytics

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Lineage data can also assist in impact analysis in case there are changes in the schema, pipelines, or business rules. From research findings, governance adoption is a challenge despite high organisational awareness. Close to 60% of the organisation either implements or develops governance practices, while only 16% have functional frameworks for governance practices. Organisational leadership, resource availability and implementation challenges continue being critical barriers (Nilsen & Bernhardsson, 2019). The results from research further reveal that poor lineage poses challenges when it comes to auditing, troubleshooting, compliance and validation of data. Consequently, the hybrid framework proposed combines lineage and metadata management and governance controls.

Enterprise Analytics and Decision-Support Capabilities

The findings indicate that lakehouses provide business intelligence support by means of dashboards, reports, and key performance indicator monitoring. Operational and strategic analysis can be performed with the help of Power BI and Tableau, for instance. Real-time analytics helps companies to analyze stream data and continuously monitor their business environment (Hasan & Akter, 2022). Solutions such as Kafka and Flink help process data faster and respond to business events in a timely manner. Moreover, lakehouse provides the capability to carry out AI and machine learning jobs with structured and unstructured enterprise data. This is an important factor that makes decision-making better as predictive analytics and models become possible.

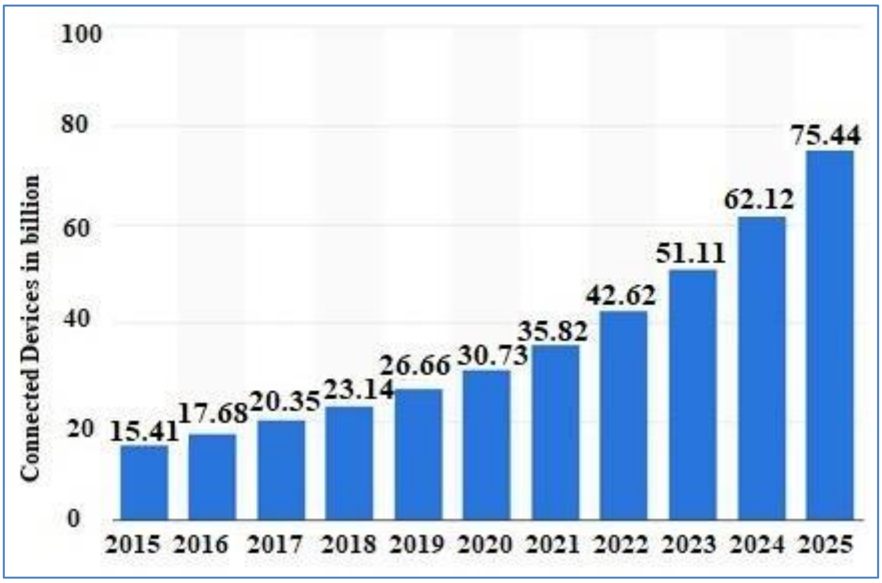

Figure 3: Internet of Things (IoT) connected devices

(Source: Alam, 2018)

Quality of decision-support increases if analytical systems use reliable data sets. Data sharing also helps to minimise inconsistencies among reports and give a consistent supply of information to the users. The accessibility of the data is improved by SQL, Python, Spark, and self-service analytics (Syed et al., 2020). Through these means, different departments have better access and can analyse the common enterprise data. Governance ensures that there is trust in analytics by ensuring that the quality, access, ownership, and proper use of the data are controlled. Lineage is used to validate the results of the analytics, and metadata assists in identifying good data sets.

TECHNICAL AND GOVERNANCE INTEGRATION CHALLENGES

Legacy System Compatibility

Many legacy systems employ out-of-date database systems and file formats as well as integration tools. Such factors lead to the emergence of obstacles when it comes to the connection of legacy systems with advanced lakehouse architecture. Data needs to be transformed prior to the process of its ingestion, thus increasing time and maintenance efforts.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Fragmented Metadata

Metadata can continue to be scattered among databases, cloud systems, applications, and departmental systems. This would make it hard for the company to determine who owns the datasets, their business meaning, quality, and how they are being used. More time could be taken by the analysts in looking for the appropriate data (Armbrust et al., 2021).

Incomplete Data Lineage

Incomplete lineage makes it hard for companies to trace how data travels from one processing stage to another. This makes it hard to debug problems, validate analytics, or understand how changes in data affect outcomes. Lineage end to end helps make these connections.

Data Duplication

Various data platforms can house various copies of the same datasets to cater to their individual needs. Such duplication will not only lead to an increased need for storage space but will also add more burden to data management (Forrow & Schiebinger, 2021). Various copies can become inconsistent with each other as well.

Real-Time Processing

New data is constantly added to the analysis environment through real-time processing. The fast movement of the data can create performance issues during the peak of the workload due to the increased processing demands. It is thus imperative that one monitors for any latencies, failures, and bottlenecks in processing.

Proposed Hybrid Lakehouse Framework

This architectural approach combines lakehouse storage, data governance, data lineage, and enterprise analytics into one system. Scalable and flexible data management is achieved with the help of cloud object storage and open table formats. A data governance layer is responsible for data ownership, data access, data quality, security, and data compliance. A metadata layer maintains information about the definitions, connections, transformations, and lineage of datasets throughout their life cycle. Processing layers provide batch and real-time processing for BI, AI, and machine learning applications. Analytical applications use governed and lineaged datasets via interfaces. Data quality monitoring detects any data issues automatically (Redyuk et al., 2021).

DISCUSSION

The implications suggest that lakehouse would be able to provide better grounds for managing data in an enterprise. The open format of storage provides flexibility, interoperability, and scalability when it comes to data processing under various workloads. At the same time, the results demonstrate that architecture alone cannot guarantee trusted enterprise analytics. Governance needs to occur on all levels, from storage and processing to metadata and analytics. Ownership is crucial here since there are problems in terms of implementing governance frameworks and the lack of organisational support (Nadal et al., 2022). These results indicate that the quality of data has an impact on the accuracy of the analysis results. Automated data validation may reveal incomplete, duplicate, inconsistent, and out-of-date data before running analysis. This makes it possible to avoid the generation of biased reports and business decisions. It is important, however, to keep performing quality checks constantly and not restrict their implementation solely to data intake operations. Another critical functionality provided within the framework of the suggested approach is data lineage. Data lineage allows users to trace data from its origin up to transformations and analytical output. In this way, users are able to ensure transparency and perform impact analysis and error identification in complex data pipelines (Rupprecht, 2020). The combination of metadata and data lineage improves this task even further. Furthermore, the outcomes reveal the positive impacts that come with sharing and governance of the datasets in enterprise analytics. For instance, business intelligence, real-time analytics, AI, and machine learning may be performed using the same datasets. This prevents duplication and enhances consistency in the outcomes from the analysis process. Nonetheless, there are other considerations that need to be taken into account for real-time workloads. In general, the outcomes suggest a hybrid model that links storage, governance, metadata, lineage, processing, and analytics (Katangoori & Katangoori, 2021). The model attempts to bridge significant gaps that exist in fragmented data settings. In this case, the model is able to prevent duplication, enhance data visibility, promote trust in the analytics, and enable quick decision-making.

LIMITATION:

In this study, secondary data are used and not primary organisational data. Hence, the success of implementation, performance measures, and organisational issues cannot be validated. Future research needs to validate the model in actual situations using actual data sets and organisations.

CONCLUSION

In summary, a combination of lakehouse architecture can be an effective solution to improving data management and analytical decision-making in enterprises. Lakehouses possess scalability in storage, flexible processing power, and various types of workloads for data analysis. Nevertheless, all these features necessitate proper governance, accurate metadata, and full data lineage. Data governance contributes to better ownership, data security, data quality, and regulation through the data lifecycle. Data lineage contributes to better visibility of the data sources, transformations, and analytical output. In turn, this paper proposes a framework that consists of all these features along with business intelligence, real-time analytics, artificial intelligence, and machine learning. This framework can help in reducing redundant data and enhancing data accessibility. There are some issues related to legacy data systems, fragmented metadata, data ownership, and complexity in implementation. Thus, it is recommended that organisations view data governance and lineage as integral parts of the lakehouse architecture.

REFERENCE LIST

[1] Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., ... & Zaharia, M. (2020). Delta Lake: high-performance ACID table storage over cloud object stores. Proceedings of the VLDB Endowment, 13(12), 3411-3424. https://www.eecs.umich.edu/courses/cse584/static_files/papers/p975-armbrust.pdf

[2] Bennouna, C., Brumbaum, H., McLay, M. M., Allaf, C., Wessells, M., & Stark, L. (2021). The role of culturally responsive social and emotional learning in supporting refugee inclusion and belonging: A thematic analysis of service provider perspectives. Plos one, 16(8), e0256743. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0256743

[3] Forrow, A., & Schiebinger, G. (2021). LineageOT is a unified framework for lineage tracing and trajectory inference. Nature communications, 12(1), 4940. https://www.nature.com/articles/s41467-021-25133-1

[4] Han, Y. S. (2021). High-Performance Data Management Architectures for Scalable Machine Learning Pipelines in Cloud Ecosystems. American International Journal of Computer Science and Technology, 3(2), 11-20. http://aijcst.org/index.php/aijcst/article/view/29

[5] Hasan, R., & Akter, S. (2022). Information system-based decision support tools: A systematic review of strategic applications in service-oriented enterprises. Review of Applied Science and Technology, 1(04), 26-65. https://rast-journal.org/index.php/RAST/article/view/55

[6] Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organising data for trustworthy Artificial Intelligence. Government Information Quarterly, 37(3), 101493. https://www.sciencedirect.com/science/article/pii/S0740624X20302719

[7] Katangoori, S., & Katangoori, A. (2021). AI-Augmented Data Governance: Enabling Intelligent Access, Lineage, and Compliance Across Hybrid Clouds. American International Journal of Computer Science and Technology, 3(6), 36-45. http://aijcst.org/index.php/aijcst/article/view/285

[8] Kheyrossadat, A. (2020). Explanation of the Qualitative Values in a Space with a view on the Capabilities of Vernacular Architecture: Case Study: Context-Based Architectural Schools. https://www.sid.ir/en/VEWSSID/J_pdf/53000120200101.pdf

[9] Nadal, S., Jovanovic, P., Bilalli, B., & Romero, O. (2022). Operationalising and automating data governance. Journal of Big Data, 9(1), 117. https://www.sciencedirect.com/science/article/pii/S1877050919305447

[10] Nilsen, P., & Bernhardsson, S. (2019). Context matters in implementation science: a scoping review of determinant frameworks that describe contextual determinants for implementation outcomes. BMC health services research, 19(1), 189. https://link.springer.com/article/10.1186/s12913-019-4015-3

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

[11] Rupprecht, L., Davis, J. C., Arnold, C., Gur, Y., & Bhagwat, D. (2020). Improving Reproducibility of Data Science Pipelines through Transparent Provenance Capture. Proc. VLDB Endow., 13(12), 3354-3368. https://www.academia.edu/download/78380173/RupprechtDavisArnoldGurBhagwat-Ursprung-VLDB20.pdf
[12] Staub, N., Haki, K., Aier, S., & Winter, R. (2022). Governance mechanisms in digital platform ecosystems: Addressing the generativity-control tension. Communications of the Association for Information Systems, 51(1), 43. https://aisel.aisnet.org/cais/vol51/iss1/43/
[13] Syed, D., Zainab, A., Ghrayeb, A., Refaat, S. S., Abu-Rub, H., & Bouhali, O. (2020). Smart grid big data analytics: Survey of technologies, techniques, and applications. Ieee Access, 9, 59564-59585. https://ieeexplore.ieee.org/abstract/document/9272794/
[14] Taherdoost, H. (2021). Data collection methods and tools for research; a step-by-step guide to choose data collection technique for academic and business research projects. International journal of academic research in management (IJARM), 10(1), 10-38. https://hal.science/Hal-03741847/
[15] Alam, T. (2018). A reliable communication framework and its use in Internet of Things (IoT). International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3(5), 450–456.https://www.researchgate.net/publication/325645304_A_Reliable_Communication_Framework_and_Its_Use_in_Internet_of_Things_IoT
[16] Visengeriyeva, L., & Abedjan, Z. (2020). Anatomy of metadata for data curation. Journal of Data and Information Quality (JDIQ), 12(3), 1-30. https://dl.acm.org/doi/abs/10.1145/3371925
[17] Ordonez, C. (2020, December). A Comparison of Data Science Systems. In International Conference on Big Data Analytics (pp. 3-11). Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-030-66665-1_1
[18] Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In Proceedings of CIDR (Vol. 8, No. 1, p. 28). sn. https://15721.courses.cs.cmu.edu/spring2023/papers/02-modern/armbrust-cidr21.pdf
[19] Redyuk, S., Kaoudi, Z., Markl, V., & Schelter, S. (2021, March). Automating Data Quality Validation for Dynamic Data Ingestion. In EDBT (pp. 61-72). https://www.ijrai.org/index.php/ijrai/article/view/341
[20] Steenbeek, I. (2023, March 20). Data management and data governance in a nutshell. Data Crossroads. https://datacrossroads.nl/2023/03/20/data-management-and-data-governance-in-a-nutshell/
[21] Subramanian, G. (2023, May 9). A practical guide to setting up your data lakehouse across AWS, Azure, GCP and Snowflake. https://www.tigeranalytics.com/perspectives/blog/a-practical-guide-to-setting-up-your-data-lakehouse-across-aws-azure-gcp-and-snowflake/