Journal of Information Systems Engineering and Management
2024, 9(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Modernizing Legacy Business Intelligence Ecosystems Through Cloud-Native Data Platforms and Automated Analytics Pipelines

Pranitha Potturi
BI architect

ARTICLEINFO

Received: 03 Apr 2024

Revised: 18 May 2024

Accepted: 29 May 2024

ABSTRACT

Many organizations still rely on the strategy of a legacy business intelligence (BI) ecosystem, characterized by a disconnected collection of servers and tools that suffer from data silos, operational latency, lack of governance, high maintenance costs, and security risks. A prolonged use of this strategy bears a strong risk of a complete system breakdown. An objective evidence-based framework is proposed for the analysis of a current-state legacy BI ecosystem, assessment of its modernization potential, and estimation of the benefits from updating to an architecture based on cloud-native data platforms with automated analytics pipelines supported by AI. Cloud-native data platforms are examined in detail, distinguishing between cloud data lakes and data warehouses, including their hybrid-coexistence, as well as assessing their scalability, resilience, cost model, vendor lock-in, and fit within multi-cloud/data sovereignty governance. Automated analytics pipelines with an end-to-end approach are considered. Roadmapping steps for the migration from the legacy BI architecture to a cloud-native one are then briefly outlined.

Keywords: Cloud-Native Data Platforms, Legacy Business Intelligence Modernization, Automated Analytics Pipelines, Cloud Data Architecture, Data Pipeline Automation, Business Intelligence Transformation, Data Lakehouse Architecture, Real-Time Data Analytics, Scalable Data Integration, Enterprise Analytics Modernization.

1. INTRODUCTION

A business intelligence (BI) ecosystem is the full set of software tools and infrastructure used to develop, run, monitor, manage, and govern BI workloads over their lifecycle, turning organizational data into the information and knowledge that decision-makers rely on. In many organizations, however, this ecosystem has grown piecemeal across successive technology cycles into a fragmented architecture with no unifying design, and it is this fragmentation, more than any single tool or vendor choice, that a modernization effort must ultimately address.

Modernization of a legacy business intelligence (BI) ecosystem is conventionally assessed through a budget mindset, whereby cost consideration drives decisions for adopting next-generation tools. This piece introduces a more informed analysis approach in the form of an objective maturity framework that centers on data organizations within a broader legacy BI ecosystem.

Recent industry shifts toward cloud architectures leverage multi-cloud data fabric concepts that minimize vendor lock-in and distribute analytics workloads closer to analytical consumers. Within this environment, cloud-native data platforms and their associated end-to-end automated analytics pipelines offer pathways for modernization with clear business impacts. Cloud-native data platforms advance scalable, resilient data management models on a consumption-based OPEX cost model, while modern cloud-native automated end-to-end analytics pipelines – both ETL and/or ELT in nature – address the challenges of aging analytics pipelines in a scalable and sustainable manner.

2. THE LEGACY BI LANDSCAPE: LIMITATIONS AND RISKS

Legacy BI ecosystems are architecturally constrained, adversely impacting performance, scalability, security, and maintenance costs. Data silos grow as horizontal scale reaches its limits, driving analysts toward shadow BI initiatives. Slow data refresh leads to governance gaps, associated compliance risk, and low confidence in decision

Journal of Information Systems Engineering and Management
2024, 9(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

making. Outdated hardware becomes a maintenance burden, while often-beleaguered IT departments struggle to meet ever-growing user demand without introducing excessive lead time or crippling organizational agility. When cloud provides an appealing new solution, vendor lock-in makes management of the BI environment, already a challenge for multinational enterprises, increasingly complex.

The cost of maintaining these fragile architectures and keeping pace with emerging user demand remains proportionately small compared to the economic risks introduced by data latency and quality — and that low proportion is set to grow. Security requirements are shifting, companies that relied on legacy BI ecosystems for a stable and secure application set face a paradox: moving to cloud is risky, but remaining behind is also becoming untenable. Combined, these pressures expose the gaping anomalies and structural weaknesses of the underlying data architecture. Slow refresh leads to unmonitored data quality, which in turn leads to unmonitored security and compliance risk. Data objects become stale and are left unattended, even as their provenance is forgotten.

TABLE I Data Lake Vs. Data Warehouse: Cloud-Native Comparison

Dimension	Data Lake	Data Warehouse
Data Structure	Raw / native format, schema-on-read	Structured, schema-on-write
Storage Cost Model	Low-cost, per-GB ingested (e.g., S3-compatible)	Higher-cost, optimized for query performance
Query Performance	Variable; needs engines (Drill, Presto)	High, purpose-built for analytics
Compute/Storage Coupling	Decoupled, scaled independently	Often coupled with storage engine
Best-Fit Workload	Exploratory, high-volume raw/ML analytics	High-volume analytical workloads

Journal of Information Systems Engineering and Management
2024, 9(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table II Migration Pattern Comparison

Pattern	Complexity	Relative Cost	Time-to-Value
Lift-and-Shift	Low	Low	Fast
Refactor	Medium	Medium	Medium
Re-platform	High	Medium-High	Medium
Re-build	Very High	High	Slow

3. CLOUD-NATIVE DATA PLATFORMS: ARCHITECTURES AND ADVANTAGES

Cloud-native characteristics—including virtualization, microservices, automated orchestration, and elastic scaling—affect data platforms in architecture and cost structure. Multi-cloud platforms avoid vendor lock-in but increase complexity, especially for governance and observability. Case studies confirm faster development of production-ready data assets.

Applications and services can be deployed in isolated containers or groups of containers in different subnets, allowing a high degree of independence. Using the services of one provider from an application located on another is possible but should be avoided whenever the platform offers local equivalents.

Data are usually stored in a combination of data lakes and data warehouses. Data lakes are low-cost, high-scale objects that can hold any type of data for an indefinite amount of time. Building data lakes with inexpensive storage hardware is usually possible, and cloud providers offer S3-compatible storage services where cost is based only on the number of ingested gigabytes and the number of read/write operations performed.

TABLE III Automated Analytics Pipeline Stages

Stage	Representative Methods	Key Consideration
Ingestion	Streaming (IoT/CDC) or Batch (ERP/RDBMS)	Latency vs. source volatility
Processing	ETL or ELT transformation	Schema evolution, data quality
Storage	Data lake and/or data warehouse	Cost vs. query performance

Journal of Information Systems Engineering and Management
2024, 9(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Stage	Representative Methods	Key Consideration
Modeling	Dimensional/semantic modeling	Consumption pattern alignment
Delivery	BI dashboards, APIs, ML features	Governance and access control

A. Data Lakes versus Data Warehouses in Cloud Environments

The concept of data lakes contrasts with that of data warehouses. Data lakes support the storage of explicit data in native formats and on native processing engines, where compute and storage are scaled independently. Data warehouses support explicit structure of the data, where used storage formats and engines provide high query performance. Data lakes are best suited for scenarios in which query performance is acceptable, while open query engines such as Apache Drill or Apache Presto may be applied. A common suggestion is to combine both concepts in a hybrid approach, using data lakes for exploratory, high-volume raw and ML-oriented analytics, while relying on data warehouses for structured, high-volume analytical workloads. Data vault considerations influence the choice of degree of normalization in data lakes.

Some cloud-native data platforms support the concept of data warehouse and data lake as part of the same service (such as Azure Synapse and Google BigQuery). In this case, the decision whether to process a given workload in the data warehouse or the data lake can be mapped to the decision of whether Elastic MapReduce (EMR) or Amazon Redshift should be used for the workload. Despite data warehouse and data lake being conceptually separate entities, many of these services blend elements of both, and in several implementations the data lake component remains less mature than the data warehouse component. These services show advantages of operating the data warehouse and data lake components as a single entity, such as cost and ease-of-use, since users only interact with one service.

Mathematical Formulas:

$$TCO_{cloud} = C_{compute} + C_{storage} + C_{transfer} + C_{ops} \quad (1)$$

where TCO_{cloud} denotes the consumption-based operating expenditure (OPEX) total cost of ownership of the cloud-native platform, aggregating compute, storage, data-transfer, and operational (DevOps/governance) cost components.

$$\Delta L = [(L_{legacy} - L_{cloud}) / L_{legacy}] \times 100\% \quad (2)$$

where ΔL is the percentage reduction in data latency achieved by migrating from the legacy BI ecosystem (L_{legacy}) to the cloud-native, automated analytics pipeline (L_{cloud}).

$$ROI = (B_{annual} - C_{annual}) / C_{annual} \quad (3)$$

where B_{annual} represents the estimated annual business benefit (e.g., faster decision-making, reduced downtime) and C_{annual} represents the annualized cost of the cloud-native architecture during the business-as-usual phase.

4. AUTOMATED ANALYTICS PIPELINES: CONCEPTS AND COMPONENTS

An end-to-end analytics pipeline encompasses the full spectrum of converting raw data into business insights. The main stages—ingestion, processing, storage, modeling, and delivery—must all be considered to avoid pitfalls. An effective analytics pipeline highlights data flow through the system and emphasizes data quality. Data lineage, including consumable formats and state changes, requires careful management, while underlying metadata has a significant effect.

Recent improvements enable faster data ingestion with lower latency, but this advantage is lost without careful planning. Business-urgent use cases must be identified, matched with suitable ingestion methods (streaming versus batch), and implemented as part of an overall data strategy, governed by the data product and data domain principles.

Journal of Information Systems Engineering and Management
2024, 9(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Choosing the correct transformation paradigm—going for ELT to leverage modern cloud-native data platform capabilities or traversing the other way for more complex processing—enables the simplification of ingestion and automation of data-model schema evolution. Although the number of orchestration tools is growing and becoming easier to use, care must be taken in their implementation: dependencies must be well understood and carefully applied in a testing environment, while scheduling control must accommodate daily and seasonal variability.

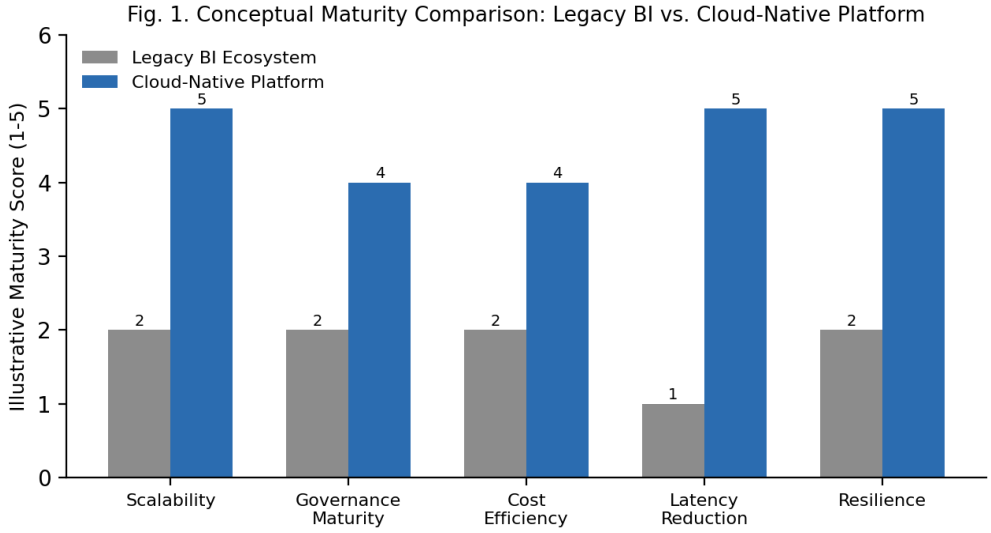

Fig. 1. Conceptual maturity comparison across five assessment dimensions (illustrative scoring, 1–5 scale).

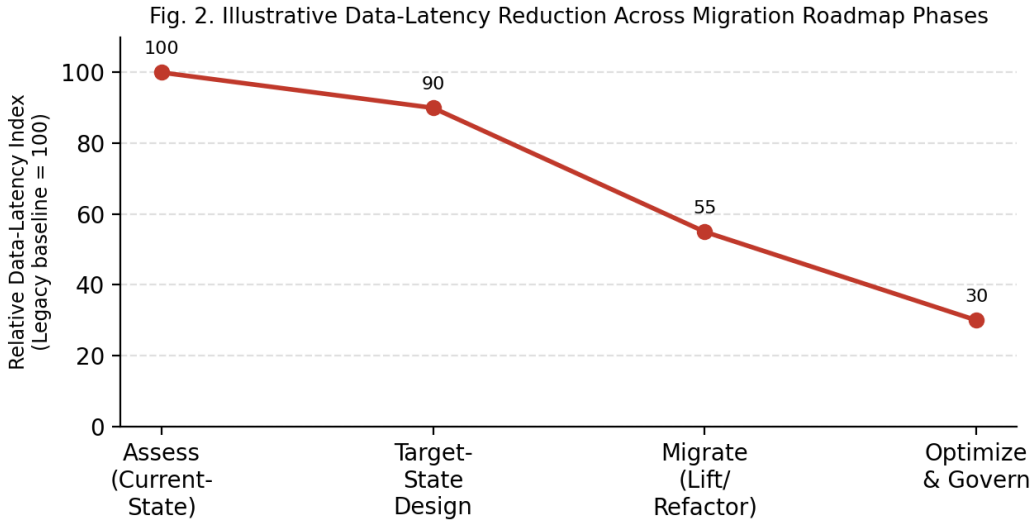

Fig. 2. Illustrative relative data-latency index across the four migration roadmap phases (legacy baseline = 100).

A. Data Ingestion, Transformation, and Orchestration

Data ingestion methods can be classified into streaming and batch categories. Streaming methods continuously load data from producers—such as IoT devices—into data platforms for real-time processing. Batch methods periodically ingest data from persistent sources (e.g., relational databases, ERP systems) that usually do not change during processing. These methods vary widely in complexity and can involve performing complex operations on the ingested data, such as cleansing, validation, normalization, and schema evolution. To facilitate correction of errors, updates,

Journal of Information Systems Engineering and Management
2024, 9(2)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

and changes in structure, extraction, loading, and transformation can also be decoupled and performed in different phases.

Automated orchestration and execution of these operations can be accomplished through dedicated tools that allow specifying the main dependencies and relationships among data sources, destinations, and functions in operations. Data pipelines typically run on a periodic basis defined by the users or data demand. Scheduling can be defined in various ways, such as by applying an event-based model to detect when the operations of particular sources or destination finish or by specifying particular points in time (e.g., the last day of each month).

5. MIGRATION STRATEGIES FROM LEGACY TO CLOUD-NATIVE ARCHITECTURES

Assessing BI modernization initiatives responds to organizational development drivers and prepares a roadmap addressing critical risks. Recent trends and cloud-native characteristics suggest cloud-native data platforms and automated analytics pipelines for the updated data engineering foundation. Assessment sets the current-state, maps to the target-state, aligns stakeholders, and defines milestones. Risks affect both technology and organization. Mitigation plans address technical concerns. Governance—covering people and processes for all data consumption, management, and engineering activities—supports enablement of self-service analytics channels. Success metrics measure foundation resilience and enable long-term cost migration and ROI estimation.

Modernization enables BI ecosystems to support evolving business requirements in response to organizational development drivers. Migration from legacy architectures to cloud-native data platforms and automated analytics pipelines mitigates stale data and labor-intensive maintenance of analytics models. For illustration, consider a hypothetical 20-plus-year legacy BI deployment spanning eight BI tools across four vendor products, with BI teams distributed across multiple departments: in such a scenario, migration would replace the decades-old data warehouse with modern cloud-native components. Business capabilities identified through an assessment framework inform mitigation of critical risks to roadmapping. Recent trends combined with cloud-native characteristics expose data-platforming and analytics-pipelining approaches for the new data-engineering foundation.

A. Assessment, Road mapping, and Risk Management

An assessment framework supports a systematic analysis, produces a roadmap for prioritizing migration initiatives, and serves as a blueprint for governance and change management. Four steps are involved. First, the current state is assessed along multiple dimensions, reflecting structured data-maturity assessment categories (e.g., technology, governance, and organizational readiness) and documenting the aspects that influence business objectives. Next, the desirable target-state is defined, and stakeholders are aligned across multiple business functions regarding the top priorities. Third, the identified initiatives are combined into a coherent milestone plan that considers logical dependencies and the availability of required skills and resources. Finally, risk management assesses both technical and organizational risks associated with the planned initiatives and prepares corresponding mitigation strategies and fallback plans.

The pattern of migration should be chosen based on the technical compatibility of the legacy and cloud-native architectures and the associated costs. The simplest step is lift-and-shift, used whenever close-to-zero latency is not a priority, the cloud provider's services for network security, data encryption, access control, and compliance can be leveraged, and recommended cost models offer savings. A more demanding option is refactoring by optimizing the legacy architecture for the cloud, while re-platforming replaces the entire ecosystem with cloud-native equivalents. Organizations that have extensive experience in cloud-native development and can afford significant investments may consider a complete re-build. If core workloads involve data-intensive learning and decision-making, cost migration and return-on-investment calculations should include the cost of the cloud-native architecture in the business-as-usual phase and quantify the benefits of the upgraded capabilities.

6. CONCLUSION

Events and trends examined in the introduction suggest that legacy BI ecosystems are neither sustainable nor likely to survive external influences. A growing body of evidence demonstrates that cloud-native data platforms and automated analytics pipelines can overcome some of the most pressing architectural constraints. Therefore, an innovative approach to modernization finds support in both qualitative and quantitative data.

The analysis of legacy refresh pathways through an objective, evidence-based framework confirms the potential of cloud-native data platforms and automated analytics pipelines to deliver demonstrably beneficial results. Migration reduces data gravity by distributing data processing closer to insight and lowers data latency through end-to-end management of analytics and machine-learning pipelines. The pathway takes organizations from a traditional three-tier architecture toward a data-lake-and-analytics-pipeline enablement. Moving to the cloud or a hybrid cloud-multi-cloud strategy reduces infrastructure responsibility and thus shifts resource allocation from operations into analysis, while aligning organizational and delivery focus to support accurate and rapid decision-making. However, the pace of external change makes 'merely' replacing the data platform insufficient: the choice of data platform should future-proof risk management and bolster control while embedding observability and traceability into the solution.

REFERENCES

[1] Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. International Journal of Information Management, 49, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

[2] Abu Rumman, A., & Al-Abbadi, L. (2023). Structural equation modeling for impact of data fabric framework on business decision-making and risk management. Cogent Business & Management, 10(2), 2215060. https://doi.org/10.1080/23311975.2023.2215060

[3] Akidau, T., Bradshaw, R., Chambers, C., Chernyak, S., Fernández-Moctezuma, R. J., Lax, R., McVeety, S., Mills, D., Perry, F., Schmidt, E., & Whittle, S. (2015). The Dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. Proceedings of the VLDB Endowment, 8(12), 1792–1803. https://doi.org/10.14778/2824032.2824076

[4] Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. Proceedings of the 11th Conference on Innovative Data Systems Research (CIDR 2021).

[5] Balalaie, A., Heydarnoori, A., & Jamshidi, P. (2016). Microservices architecture enables DevOps: Migration to a cloud-native architecture. IEEE Software, 33(3), 42–52. https://doi.org/10.1109/MS.2016.64

[6] Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), 1–52. https://doi.org/10.1145/1541880.1541883

[7] Becker, J., Knackstedt, R., & Pöppelbuß, J. (2009). Developing maturity models for IT management: A procedure model and its application. Business & Information Systems Engineering, 1(3), 213–222. https://doi.org/10.1007/s12599-009-0044-5

[8] Chaudhuri, S., Dayal, U., & Narasayya, V. (2011). An overview of business intelligence technology. Communications of the ACM, 54(8), 88–98. https://doi.org/10.1145/1978542.1978562

[9] Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165–1188. https://doi.org/10.2307/41703503

[10] Demirkan, H., & Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. Decision Support Systems, 55(1), 412–421. https://doi.org/10.1016/j.dss.2012.05.048

[11] Dhaouadi, A., Bousselmi, K., Gammoudi, M. M., Monnet, S., & Hammoudi, S. (2022). Data warehousing process modeling from classical approaches to new trends: Main features and comparisons. Data, 7(8), 113. https://doi.org/10.3390/data7080113

[12] Fahmideh, M., Low, G., Beydoun, G., & Daneshgar, F. (2016). Cloud migration process—A survey, evaluation framework, and open challenges. Journal of Systems and Software, 120, 31–69. https://doi.org/10.1016/j.jss.2016.06.068

[13] Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2020). A zone reference model for enterprise-grade data lake management. In 2020 IEEE 24th International Enterprise Distributed Object Computing Conference (pp. 57–66). IEEE. https://doi.org/10.1109/EDOC49727.2020.00017

[14] Grozev, N., & Buyya, R. (2014). Inter-cloud architectures and application brokering: Taxonomy and survey. Software: Practice and Experience, 44(3), 369–390. https://doi.org/10.1002/spe.2168

[15] Hai, R., Koutras, C., Quix, C., & Jarke, M. (2023). Data lakes: A survey of functions and systems. IEEE Transactions on Knowledge and Data Engineering, 35(12), 12571–12590. https://doi.org/10.1109/TKDE.2023.3270101

[16] Hasan, M. H., Osman, M. H., Admodisastro, N. I., & Muhammad, M. S. (2023). Legacy systems to cloud migration: A review from the architectural perspective. Journal of Systems and Software, 202, 111702. https://doi.org/10.1016/j.jss.2023.111702

[17] Hasan, M. R., & Legner, C. (2023). Understanding data products: Motivations, definition, and categories. In Proceedings of the 31st European Conference on Information Systems. AIS Electronic Library

[18] Herschel, M., Diestelkämper, R., & Ben Lahmar, H. (2017). A survey on provenance: What for? What form? What from? The VLDB Journal, 26(6), 881–906. https://doi.org/10.1007/s00778-017-0486-1

[19] Khajeh-Hosseini, A., Greenwood, D., Smith, J. W., & Sommerville, I. (2012). The cloud adoption toolkit: Supporting cloud adoption decisions in the enterprise. Software: Practice and Experience, 42(4), 447–465. https://doi.org/10.1002/spe.1072

[20] Khatri, V., & Brown, C. V. (2010). Designing data governance. Communications of the ACM, 53(1), 148–152. https://doi.org/10.1145/1629175.1629210

[21] Machado, I. A., Costa, C., & Santos, M. Y. (2022). Data mesh: Concepts and principles of a paradigm shift in data architectures. Procedia Computer Science, 196, 263–271. https://doi.org/10.1016/j.procs.2021.12.013

[22] Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., & Ghalsasi, A. (2011). Cloud computing—The business perspective. Decision Support Systems, 51(1), 176–189. https://doi.org/10.1016/j.dss.2010.12.006

[23] Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., & Arocena, P. C. (2019). Data lake management: Challenges and opportunities. Proceedings of the VLDB Endowment, 12(12), 1986–1989. https://doi.org/10.14778/3352063.3352116

[24] Khoda Parast, F., Sindhav, C., Nikam, S., Yekta, H. I., Kent, K. B., & Hakak, S. (2022). Cloud computing security: A survey of service-based models. Computers & Security, 114, 102580. https://doi.org/10.1016/j.cose.2021.102580

[25] Petcu, D. (2013). Multi-cloud: Expectations and current approaches. In Proceedings of the 2013 International Workshop on Multi-Cloud Applications and Federated Clouds (pp. 1–6). ACM. https://doi.org/10.1145/2462326.2462328

[26] Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. Journal of Intelligent Information Systems, 56(1), 97–120. https://doi.org/10.1007/s10844-020-00608-7

[27] Stonebraker, M., Çetintemel, U., & Zdonik, S. (2005). The eight requirements of real-time stream processing. ACM SIGMOD Record, 34(4), 42–47. https://doi.org/10.1145/1107499.1107504

[28] Trieu, V. H. (2017). Getting value from business intelligence systems: A review and research agenda. Decision Support Systems, 93, 111–124. https://doi.org/10.1016/j.dss.2016.09.019

[29] Vassiliadis, P. (2009). A survey of extract-transform-load technology. International Journal of Data Warehousing and Mining, 5(3), 1–27. https://doi.org/10.4018/jdwm.2009070101

[30] Wider, A., Verma, S., & Akhtar, A. (2023). Decentralized data governance as part of a data mesh platform: Concepts and approaches. In 2023 IEEE International Conference on Web Services (pp. 746–754). IEEE. https://doi.org/10.1109/ICWS60048.2023.00101

[31] Golfarelli, M., Rizzi, S., & Cella, I. (2004). Beyond data warehousing: What's next in business intelligence? Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP, 1–6.