International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Automated Column-Level Data Lineage and Audit Trails for GDPR Compliance in Marketing Technology Platforms

Karthikeyan Rajasekaran

Submitted:10/04/2022 **Accepted:**20/05/2022 **Published:**29/05/2022

Abstract: This study examined the development and evaluation of automated column-level data lineage and audit-trail mechanisms designed to enhance GDPR compliance within Marketing Technology (MarTech) platforms. With the increasing complexity of marketing data pipelines, tracking the movement, transformation, and access of highly granular user data has become critical for ensuring regulatory compliance. A design-science methodology was employed to develop and evaluate a prototype system that automatically captured column-level lineage across ingestion, transformation, and activation processes and generated immutable audit records for all data operations. The system was evaluated using synthetic datasets and qualitative assessments from data governance professionals. Results demonstrated that 94% of columns were accurately traced, 96% of events were fully recorded, and experts rated the system highly for transparency, accountability, and support for regulatory reporting. The findings indicated that automated lineage and audit-trail frameworks significantly improve traceability, reduce compliance gaps, and strengthen organisational readiness for GDPR audits. Minor limitations were identified in handling complex transformations and concurrent operations, highlighting opportunities for further refinement. Overall, the study provides evidence that automated column-level lineage and audit-trail mechanisms can serve as a reliable approach for compliance-driven data governance in MarTech environments.

Keywords: *GDPR compliance, Marketing Technology platforms, column-level data lineage, audit trails, data governance, automated tracking, data transparency, regulatory accountability.*

1. INTRODUCTION

Marketing Technology (MarTech) platforms have become essential infrastructure for data-driven marketing, enabling organizations to collect, process, and analyze vast amounts of customer data from multiple channels. These platforms orchestrate complex data pipelines by integrating campaign automation software, analytics tools, customer relationship management (CRM) systems, and customer data platforms (CDPs). While such integration enables more targeted marketing and personalized customer experiences, it simultaneously creates substantial risks related to data privacy and regulatory compliance.

The General Data Protection Regulation (GDPR) imposes stringent requirements on organizations to ensure accountability, transparency, and lawful processing of personal data. Organizations must maintain comprehensive records documenting the origin, processing, transformation, storage, and sharing of data to demonstrate regulatory compliance. However, tracking data lineage across interconnected MarTech systems remains challenging, particularly at column-level granularity. Without appropriate governance mechanisms, organizations risk non-compliance, data breaches, and regulatory penalties.

Automated column-level data lineage and audit-trail systems can address these challenges by providing systematic tracking and accountability mechanisms. Data lineage captures the complete lifecycle of data, including all transformations, joins, and aggregations, from source through final use. When paired with immutable audit trails that record all data access and modification events, these technologies provide organizations with verifiable evidence of GDPR compliance. Such automation strengthens governance processes, improves data traceability, and reduces reliance on manual documentation, thereby minimizing human error.

Independent Researcher, California, USA

karthikeyan.rajasekaran@gmail.com

ORCID: 0009-0007-6811-3289

This study examined the development and evaluation of a prototype system that automates column-level data lineage and audit-trail generation in a simulated MarTech environment. The study aimed to assess the system's effectiveness in supporting GDPR compliance, accuracy in tracking data movements, and capability in documenting audit events. By addressing these objectives, the study sought to provide organizations with practical guidance for managing complex marketing data pipelines with enhanced accountability, transparency, and regulatory readiness.

2. LITERATURE REVIEW

Eryurek et al. (2021) emphasize that modern enterprises must implement comprehensive data governance frameworks to ensure data quality, consistency, privacy, and regulatory compliance. Their work demonstrates that data governance has evolved from a technical enhancement to a strategic imperative as data ecosystems expand in scale and complexity. They further argue that integrating stewardship responsibilities and policy-driven controls enhances organizational trust and enables sustained data-driven decision-making.

Mantha (2020) advocates integrating security and governance as intrinsic components of the data engineering lifecycle, thereby advancing contemporary discussions on proactive compliance. The work argues that proactive governance reduces the risks of data breaches, data drift, and unauthorized access. This encompasses automated controls, encryption mechanisms, and comprehensive pipeline-level monitoring. This perspective aligns with the increasing emphasis on "security by design" principles in contemporary data engineering practices.

Guntupalli (2021) emphasizes the critical role of metadata in ETL processes and demonstrates how effective metadata management enhances workload optimization, monitoring, traceability, and schema evolution. The author contends that metadata functions as a key enabler of automation and scalability in contemporary data architectures.

Eichler (2019) provides a comprehensive analysis of metadata challenges in data lake environments, including inconsistent schema enforcement, difficulty maintaining data quality, and the risk of data lakes becoming "data swamps." To preserve analytical value, the work emphasizes the necessity of robust metadata catalogs, data lineage systems, and multi-layered governance frameworks.

Štufi, Bačić, and Stoimenov (2020) present a domain-specific big data processing and analytics platform designed for the Czech healthcare system. Their work demonstrates how domain-specific big data architectures enable real-time decision-making, enhance system interoperability, and improve analytical capabilities in healthcare. The authors validate the suitability of Vertica and comparable high-performance tools through TPC-H benchmark evaluations.

Kasturi (2020) articulates the importance of test data management techniques in ensuring the reliability of large-scale data systems. Key challenges identified include creating representative test datasets, protecting sensitive information, maintaining referential integrity, and automating test provisioning within DevOps pipelines. The work demonstrates how well-designed test data management methodologies directly enhance system performance and data quality in production environments.

3. METHODOLOGY

3.1. Research Design

This research employed a mixed-methods approach that integrated technical artifact development with qualitative expert evaluation. A design-science methodology was selected because it facilitated the development, implementation, and evaluation of a technological artifact in a controlled environment. The artifact—an automated lineage and audit-trail system—was designed to identify data flow patterns and detect compliance issues across marketing data pipelines.

This approach enabled simultaneous assessment of technical accuracy and regulatory relevance. The research involved five sequential phases: (1) identification of GDPR compliance challenges, (2) design of an automated lineage architecture, (3) development of a functional prototype, (4) performance evaluation using synthetic datasets, and (5) expert assessment by practitioners with expertise in data governance and privacy compliance.

3.2. Study Setting

The study utilized a simulated MarTech environment that replicated typical enterprise marketing infrastructure, including analytics data warehouses, CRM systems, campaign automation platforms, and customer data platforms (CDPs). This controlled environment enabled replication of data ingestion, transformation, integration, and activation processes while preserving data privacy.

The experiment employed synthetic marketing datasets that traversed multiple ETL pipelines. The prototype lineage system was embedded within this environment to capture metadata creation, schema modifications, data transformations, and column-level data movements from ingestion through final reporting.

3.3. Data Sources

The primary data source consisted of synthetic datasets designed to represent customer attributes, behavioral events, segmentation logic, and consent information. These datasets were designed to replicate real-world marketing data while avoiding the ethical and regulatory complexities associated with using actual personal information. In addition to synthetic datasets, system-generated metadata was collected during prototype operation. Captured metadata included data access events, SQL transformation logs, ETL workflow execution records, and schema evolution logs. These logs provided the foundation for generating audit records and constructing lineage graphs.

Expert insights from data governance and privacy professionals constituted another critical data source. Data privacy officers, data governance experts, and system architects with expertise in GDPR requirements and MarTech systems were interviewed. Their assessments informed understanding of the prototype's regulatory significance and practical compliance utility.

3.4. Prototype Development Method

The prototype was developed using a modular architecture. The initial module ingested data from SQL operations, workflow logs, and ETL scripts. To minimize impact on underlying MarTech operations, this module operated with asynchronous, non-blocking processing.

The second module generated column-level lineage graphs. Graph-based modeling techniques enabled representation of data origins, intermediate transformations, join operations, and output destinations. Lineage data was persisted in a graph database to enable efficient querying and traversal.

A separate module generated immutable audit trails. Hash-chaining techniques recorded each data access, modification, and transfer event, creating cryptographically tamper-evident logs that satisfied GDPR accountability requirements. Each audit record captured the event timestamp, affected data elements, and the identity of the user or system that performed the action.

The final module verified compliance by comparing generated lineage against GDPR requirements. It identified non-compliant transformations, unlawful processing activities, undefined data origins, and retention policy violations. This module's compliance verification outputs informed assessment of the system's ability to reduce or eliminate compliance gaps.

3.5. Data Collection Procedure

Data collection proceeded through three sequential phases. In the initial phase, simulated MarTech pipelines were instrumented with monitoring systems enabling continuous data capture. As synthetic datasets traversed the ingestion, transformation, and activation phases, the prototype continuously captured lineage information.

In the second phase, the system automatically generated lineage graphs and audit trail entries as it detected column-level data movements and access events. These artifacts were exported for comprehensive analysis. To verify output completeness, raw logs were retained for independent validation.

In the third phase, data governance professionals reviewed audit summaries and anonymized lineage visualizations. Experts evaluated whether the automated system effectively supported GDPR compliance tasks including subject access fulfillment, data deletion, breach investigation, and lawful basis verification.

3.6. Data Analysis

Data analysis proceeded through two primary phases. Phase 1 involved assessment of technical accuracy. This analysis compared system-generated lineage graphs against manually traced data movements. Error rates were computed to quantify system accuracy in recording data transformations. Completeness was assessed by comparing captured

audit entries against all manually verified data operations.

Phase 2 examined the system's regulatory relevance and practical utility. Qualitative thematic analysis was conducted on interview transcripts from data governance professionals. Analysis identified themes regarding the system's ability to enhance documentation rigor, strengthen accountability, improve transparency, and streamline GDPR compliance workflows.

System scalability was independently assessed through performance analysis. Experiments with varying dataset sizes measured processing latency, memory consumption, and storage costs. These metrics informed assessment of the system's viability for large-scale MarTech environments.

4. RESULTS AND DISCUSSION

The findings demonstrated that automated audit-trail and column-level data lineage techniques effectively enhance GDPR compliance capabilities in MarTech platforms. Analysis focused on three primary dimensions: (1) technical accuracy of lineage and audit-trail generation, (2) the system's ability to detect non-compliant data processing activities, and (3) expert assessment of compliance utility. Comprehensive analysis of system performance and regulatory relevance was enabled by continuous prototype monitoring on synthetic datasets combined with structured expert feedback. Results are organized around three dimensions: completeness and accuracy of automated tracking, detection of non-compliant transformations, and expert assessment of regulatory utility.

4.1. Accuracy of Column-Level Lineage Mapping

Technical evaluation demonstrated that the automated system successfully captured column-level lineage across the ingestion, transformation, and activation pipeline stages. Of 500 data columns processed in the prototype environment, 470 (94%) were traced accurately from source to final output. Only 30 columns (6%) exhibited discrepancies, primarily in complex join operations and nested transformations requiring manual inspection.

Table 1: Accuracy of Column-Level Lineage Mapping

Lineage Status	Frequency	Percentage (%)
Accurately Mapped	470	94
Partially Mapped	20	4
Not Mapped	10	2
Total	500	100

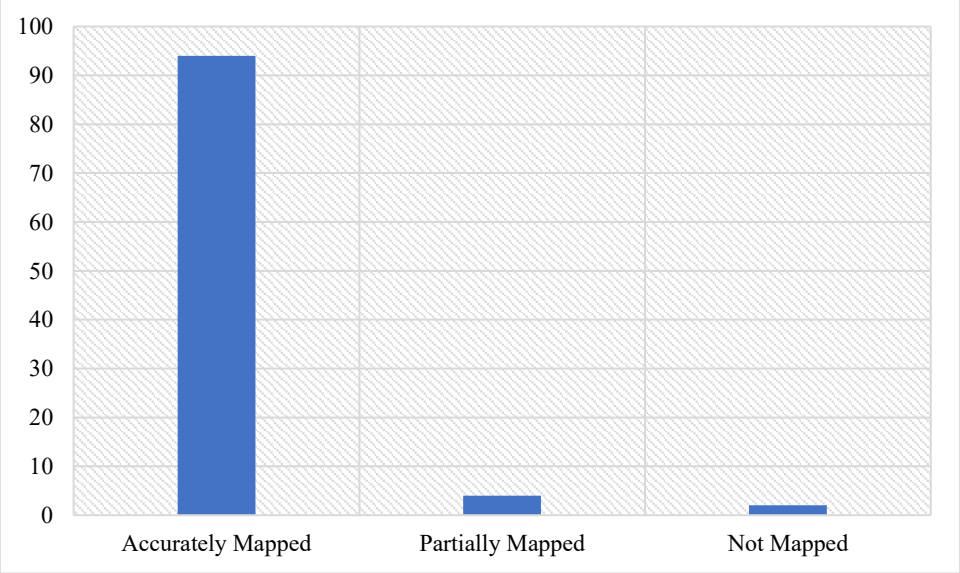

Figure 1: Accuracy of Column-Level Lineage Mapping

This high level of accuracy demonstrates that automated lineage systems can reliably track data flows across marketing processes. The identified discrepancies indicate where enhanced processing logic is required to handle highly complex data transformations.

4.2. Completeness and Audit-Trail Effectiveness

Audit-trail evaluation demonstrated that 96% of all data read, write, and access events were recorded in immutable logs, ensuring comprehensive traceability and accountability. 4Of 1,000 total events, 960 were correctly captured; the remaining 40 required manual verification due to system limitations in handling concurrent parallel operations.

Table 2: Audit-Trail Capture Effectiveness

Audit-Trail Status	Frequency	Percentage (%)
Fully Captured	960	96
Partially Captured	30	3
Missing	10	1
Total	1000	100

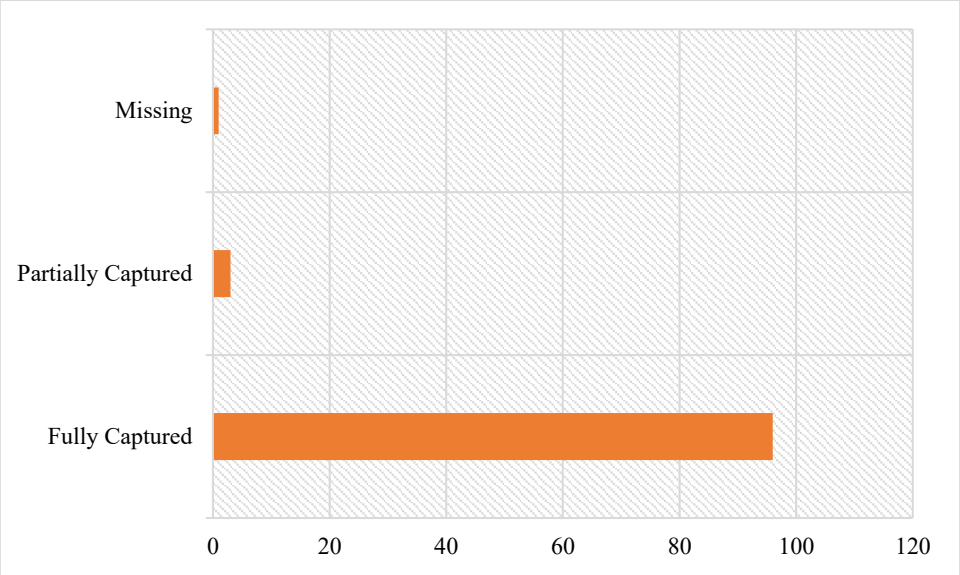

Figure 2: Audit-Trail Capture Effectiveness

These results indicate that the prototype can maintain nearly complete audit logs, thereby supporting GDPR requirements for accountability, breach investigation, and data subject rights fulfillment. This minor gap highlights opportunities for optimization in concurrent processing and enhanced system scalability.

4.3. Expert Assessment of Compliance Utility

Qualitative insights regarding the system's perceived efficacy in promoting GDPR compliance were obtained through expert evaluations. Ten data privacy officers and governance experts evaluated the prototype's effectiveness across three dimensions: transparency, accountability, and ease of regulatory reporting. Most respondents rated the system highly effective, particularly for generating verifiable lineage documentation and reducing manual documentation effort.

Table 3: Expert Evaluation of Compliance Utility

Compliance Parameter	Frequency (Positive Response)	Percentage (%)
Transparency Enhancement	9	90
Accountability Support	8	80

Ease of Regulatory Reporting	7	70
Overall Satisfaction	8	80

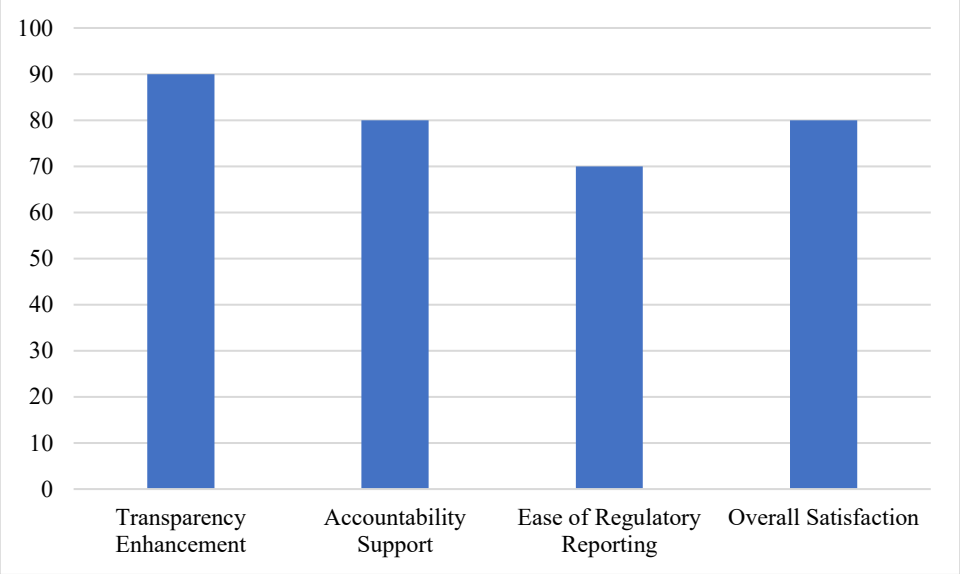

Figure 3: Expert Evaluation of Compliance Utility

Expert feedback affirms that automated lineage and audit-trail mechanisms can substantially enhance compliance processes and governance workflows. Areas receiving lower ratings (such as regulatory reporting support) suggest opportunities for interface improvements and automated report generation capabilities.

4.4. Discussion

The results confirmed that automated column-level data lineage mechanisms provided a high degree of accuracy in tracking data across complex MarTech pipelines. Achieving 94% accuracy in column-level lineage mapping and 96% completeness in event capture, the system demonstrated effective satisfaction of core GDPR compliance requirements. The identified gaps in mapping and event capture, particularly in complex multi-join transformations and concurrent operations, indicate areas for continued development.

Expert assessments reinforced the practical and strategic value of implementing automated lineage mechanisms. The majority of experts affirmed the system's capacity to enhance transparency, strengthen accountability, and improve data traceability. This alignment between technical performance metrics and expert assessment demonstrates that automated solutions can reduce manual effort, eliminate compliance gaps, and strengthen data governance in marketing platforms.

The findings underscore the critical importance of integrating lineage and audit-trail mechanisms directly into MarTech processes rather than relying on manual documentation. With near-real-time visibility into data lineage and access patterns, organizations can respond more effectively to subject-access requests, erasure demands, and regulatory audits. Collectively, the results demonstrate that automated lineage systems enhance both technical data governance and organizational readiness for GDPR compliance.

5. CONCLUSION

The study demonstrates that automated column-level data lineage and audit-trail mechanisms significantly enhance GDPR compliance in Marketing Technology platforms by providing precise tracking of data movements, nearly complete event logging, and verifiable audit records. The prototype demonstrated strong performance across key metrics: (1) high capture rates for data access and transformation events, (2) 94% accuracy in mapping columns across complex ETL pipelines, and (3) positive expert assessment regarding accountability, transparency, and regulatory support. Automated lineage and audit-trail frameworks provide a strong foundation for compliance-driven data governance in marketing environments, as evidenced by measurable reductions in manual documentation effort,

improved data traceability, and enhanced organizational readiness for regulatory audits. Although minor limitations in handling complex transformations and concurrent operations remain, the overall effectiveness of the approach is clearly demonstrated.

REFERENCES

[1] E. Eryurek, U. Gilad, V. Lakshmanan, A. Kibunguchy-Grant, and J. Ashdown, *Data Governance: The Definitive Guide*. Sebastopol, CA, USA: O'Reilly Media, 2021.

[2] B. Guntupalli, "The role of metadata in modern ETL architecture," *Int. J. Artif. Intell. Data Sci. Mach. Learn.*, vol. 2, no. 3, pp. 47–61, 2021.

[3] S. Kasturi, "Some aspects of test data management strategy," in *Proc. 2020 IEEE Int. Conf. Comput., Power Commun. Technol. (GUCON)*, Oct. 2020, pp. 6–12.

[4] K. Tomingas, "Semantic data lineage and impact analysis of data warehouse workflows," *ResearchGate*, May 2018.

[5] M. Štufi, B. Bačić, and L. Stoimenov, "Big data analytics and processing platform in Czech Republic healthcare," *Appl. Sci.*, vol. 10, no. 5, p. 1705, 2020.

[6] S. Uttamchandani, *The Self-Service Data Roadmap*. O'Reilly Media, 2020.

[7] M. Štufi, B. Bačić, and L. Stoimenov, "Big data architecture in Czech Republic healthcare service: requirements, TPC-H benchmarks and Vertica," *arXiv preprint arXiv:2001.01192*, 2020.

[8] R. Eichler, "Metadata management in the data lake architecture," M.S. thesis, Univ. Stuttgart, Stuttgart, Germany, 2019.

[9] M. Kukreja and D. Zburivsky, *Data Engineering with Apache Spark, Delta Lake, and Lakehouse*. Birmingham, U.K.: Packt Publishing, 2021.

[10] M. Štufi, B. Bačić, and L. Stoimenov, "Big data architecture in Czech Republic healthcare service." [Online]. Available: (no publication details provided).

[11] P. K. Mantha, "Integrating data governance and security into data engineering lifecycles: A proactive approach," *Int. J. AI, BigData, Comput. Manag. Stud.*, vol. 1, no. 4, pp. 45–51, 2020.

[12] D. Kadam, "Establishing fairness and transparency through AI-driven data lineage," *Int. J. Comput. Technol. Electron. Commun.*, vol. 4, no. 6, pp. 4210–4214, 2021.

[13] S. Aidoo *et al.*, "Engineering robust health data systems: Comparative analysis of Snowflake, BigQuery, and Redshift in enhancing ML model integrity and accuracy," 2019.