Automated Data Governance and Compliance Monitoring using AI & Big Data

Ujjawal Nayak

Software Development Manager
Experian Information Solutions, Inc.
Costa Mesa, CA, USA.

Abstract:
Global privacy mandates—GDPR, CCPA, FCRA, HIPAA—now cover petabyte-scale, multi-cloud data estates. Spreadsheets and quarterly audits can no longer track schema drift or regional rule changes. This paper presents an AI-driven reference architecture that automates data discovery, classification, policy-as-code enforcement, and continuous controls monitoring (CCM). A real-world marketing analytics case study shows a 40 % cut in audit-preparation hours, a 35 % fall in policy violations, and a 53 % drop in the mean-time-to-remediate, reducing data-warehouse costs by 28 % after a targeted migration and optimization program.

Keywords: data governance, compliance monitoring, AI, big data, policy-as-code, metadata catalog, continuous controls monitoring.

I. INTRODUCTION

The 2025 *AI & Compliance Market Study* reports that 52 % of organizations now rely on AI tooling for at least one compliance workflow, yet only 9 % have achieved end-to-end automation [4]. Simultaneously, a benchmark survey finds that 41 % of firms are reassessing privacy controls to counter AI-enabled attacks [7]. To satisfy regulators and repel emerging threats, governance must shift from episodic audits to continuous, evidence-based controls. Enterprises can enforce policy decisions in real time by uniting AI classifiers with distributed big-data frameworks such as Spark, Kafka, and Snowflake.

II. RELATED WORK

Early metadata catalogs captured static technical lineage. Contemporary research embeds machine-learning classifiers that auto-tag personal and sensitive attributes, feeding masking or access-control engines [8]. Open Policy Agent (OPA) popularized declarative *policy-as-code*; its Rego language unifies rule evaluation across APIs, data platforms, and CI/CD pipelines [5], while implementation guides highlight infrastructure-governance patterns [6].

III. REFERENCE ARCHITECTURE

1. **Ingestion & Lineage** – Batch (Spark, Snowflake *COPY*) and streaming (Kafka, Kinesis) jobs emit Open Lineage events.
2. **AI Classification Engine** – BERT-based NER models, plus gradient-boosted trees for structured fields, label PII (e.g., *SSN, card number*), and write tags to a central catalog.
3. **Policy-as-Code Service** – Rego rules compare tags to regional mandates (e.g., “mask for GDPR”, “retain 7 years for FCRA”) at query time [5], mapping decisions to Snowflake *TAG-BASED MASKING* or Spark UDFs.
4. **Continuous Controls Monitoring** – Prometheus scrapes policy-decision metrics; Grafana overlays real-time dashboards with historical trends in AWS Timestream.
5. **Alerting & Remediation Bots** – Airflow DAGs or Step Functions quarantine non-compliant tables and open JIRA tickets automatically.

IV. AI TECHNIQUES

- **Supervised learning** – Boosted-tree models detect structured PII; transformer-based NER tags unstructured text.
- **Unsupervised drift detection** – Auto-encoders flag schema anomalies that hint at shadow IT pipelines.
- **Reinforcement learning** – Agents tune masking thresholds to minimize false positives while maintaining zero violations: Reuters notes growing interest in such self-adjusting models [10].

V. CASE STUDY: AN ANONYMIZED ANALYTICS PLATFORM

An analytics platform processed > 3 TB/day. Guided by a Snowflake migration framework [2] and a cost-optimization playbook [3], engineers shifted workloads to Spark-on-EMR and Snowflake, then deployed the AI governance stack.

Metric	Before	After	Difference %
Audit-prep hours/quarter	120 h	72 h	–40 %
Policy violations/month	23	15	–35 %
Mean-time-to-remediate	6 h	2.8 h	–53 %
Warehouse cost (annual)	baseline	–28 %	

Key enablers included tag-based masking, error remediation bots, and unified observability. Lessons learned were later distilled into an ETL-pipeline design pattern [1].

VI. IMPLEMENTATION BEST PRACTICES

- **Catalogue everything early.** Automated classification is only as good as the underlying lineage.
- **Adopt policy-as-code from day one.** Declarative rules simplify audits and CI/CD promotion [6].
- **Instrument CCM metrics.** Track precision, recall, and mean time-between-violations; retrain models quarterly.
- **Default to privacy-by-design.** Mask by default; grant unmasked access via time-bound tokens.

VII. CHALLENGES & FUTURE WORK

Training-set bias can overmask minority identifiers, reducing analytics utility. Synthetic data generation and federated learning are promising mitigation paths [9]. Meanwhile, regulators signal interest in *self-healing* policy agents capable of recompiling rules as laws evolve.

VIII. CONCLUSION

AI-augmented governance aligns data platform velocity with regulatory change. By coupling scalable big-data tooling with intelligent policy engines, organizations achieve continuous compliance, slash audit costs, and resolve incidents faster without sacrificing analytical agility.

REFERENCES:

[1] U. Nayak, “Building a Scalable ETL Pipeline with Apache Spark, Airflow, and Snowflake,” *Int. J. Innovative Res. Creative Technol.*, vol. 11, no. 2, pp. 1-7, 2025.

[2] U. Nayak, “Migrating Legacy Data Warehouses to Snowflake,” *Int. J. Science & Technol.*, vol. 16, no. 1, pp. 1-5, 2025.

[3] U. Nayak, “Cost Optimization Strategies in Cloud Data Warehousing: A Comparative Study of AWS Redshift and Snowflake,” *Int. J. Core Eng. & Manag.*, vol. 8, no. 2, pp. 1-4, 2025.

[4] Star Compliance, *AI & Compliance 2025 Market Study*, white paper, 2025. [Online]. Available: https://www.starcompliance.com/resource/ai-in-compliance-market-study/.

[5] Open Policy Agent, “Official Documentation,” accessed Jul. 5, 2025. [Online]. Available: https://openpolicyagent.org/docs/.

[6] env0, "How Policy-as-Code Enhances Infrastructure Governance with Open Policy Agent (OPA)," env0 Blog, 2024. [Online]. Available: https://www.env0.com/blog/how-policy-as-code-enhances-infrastructure-governance-with-open-policy-agent-opa.

[7] Gallagher Re, *The 2025 Attitudes to AI Adoption and Risk Benchmarking Survey*, 2025. [Online]. Available: https://www.ajg.com/gallagherre/news-and-insights/features/2025-attitudes-to-ai-adoption-and-risk-benchmarking-survey/.

[8] Atlan, "Data Access Controls for Sensitive Financial Information," Atlan Blog, Jun. 2025. [Online]. Available: https://atlan.com/know/data-access-controls-for-sensitive-financial-information/.

[9] Atlan, "9 Financial Data Compliance Challenges to Tackle in 2025," Atlan Blog, Jul. 2025. [Online]. Available: https://atlan.com/know/data-governance/financial-data-compliance-challenges/.

[10] C. Elson, "AI agents: greater capabilities and enhanced risks," *Reuters*, Apr. 22, 2025. [Online]. Available: https://www.reuters.com/legal/legalindustry/ai-agents-greater-capabilities-enhanced-risks-2025-04-22/.