Chapter 5
Developing Sustainable Data Retention Policies:
A Machine Learning Approach to Intelligent Data Lifecycle Management

Swathi Chundru
https://orcid.org/0009-0001-3959-4825
Motivity Labs Inc., USA

Lakshmi Narasimha Raju Mudunuri
https://orcid.org/0009-0004-7945-0668
Valero Energy Corporation, USA

ABSTRACT

In the era of rapidly growing data, organizations face increasing challenges in managing vast volumes of information while adhering to regulatory requirements and optimizing storage costs. Developing sustainable data retention policies is crucial for efficient data management and ensuring compliance with privacy laws and industry regulations. This paper explores a machine learning (ML)-driven approach to intelligent data lifecycle management, offering a framework for creating sustainable data retention policies. By leveraging ML algorithms, the proposed system automates the classification, retention, and deletion of data based on predefined business rules, compliance standards, and usage patterns. This approach enhances data governance, reduces operational costs, and improves the efficiency of data management practices.

1. INTRODUCTION

1.1 Overview of Data Retention Policies

Data retention policies are essential frameworks that define how long data should be stored, when it should be archived, and when it should be deleted. These policies are driven by legal, regulatory, and operational requirements, ensuring that data is kept for as long as necessary and deleted or archived when no longer useful. Proper data retention policies help organizations manage data efficiently, mitigate

DOI: 10.4018/979-8-3693-9750-3.ch005

risk, and comply with regulations like GDPR, HIPAA, and various industry standards. These policies typically cover the classification of data, the handling of sensitive or personally identifiable information (PII), and the processes of storing, accessing, and deleting data.

In the digital era, where data is continuously generated and stored in massive quantities, organizations face the challenge of creating retention policies that balance between compliance, security, cost-effectiveness, and operational efficiency. Poor data retention practices can result in data breaches, costly storage solutions, and violations of regulations. Therefore, defining a clear and effective data retention policy is vital for long-term business success. The development of sustainable data retention policies and intelligent data lifecycle management has become increasingly important as organizations face growing volumes of data. Machine learning (ML) is emerging as a powerful tool for optimizing these policies, enabling efficient data retention strategies across diverse systems, particularly in cloud-based and multi-cloud environments (Baker & Smith, 2022; Chen & Li, 2023). By employing ML techniques, businesses can automate data retention processes, ensuring compliance with regulations while reducing the risks associated with over-retention or loss of critical information (Davis & Wright, 2023).

Recent studies emphasize the role of predictive analytics and intelligent frameworks powered by machine learning in enhancing data lifecycle management (Clark & Patel, 2024; George & Lee, 2022). These ML-driven systems offer proactive management solutions that anticipate data access needs and retention requirements, thus improving operational efficiency and sustainability (Harrison & Williams, 2023). Such techniques are especially useful in the context of distributed systems and cloud environments, where data storage and retrieval practices can be optimized dynamically (Kumar & Singh, 2024).

Sustainable data retention policies also leverage advances in deep learning and artificial intelligence (AI) to predict data usage patterns and apply retention policies accordingly (Ferris & Gupta, 2024; Lee & Sharma, 2023). This trend reflects a broader push toward intelligent infrastructure that not only retains data but also manages it efficiently, aligning with enterprise needs and regulatory frameworks (Whig, Kouser, Bhatia, Nadikattu, & Alkali, 2024). Moreover, the integration of machine learning in these areas helps address the growing challenges of data overload and compliance management (Whig, Kautish, Nadikattu, & Alkali, 2024).

The role of machine learning extends beyond simple automation to more sophisticated tasks such as enhancing predictive maintenance in supply chain systems (Koushik, Mittal, Jain, & Whig, 2024). These innovations also support the design of intelligent sustainable infrastructures for procurement and distribution, contributing to the broader goal of creating a data-driven, sustainable enterprise ecosystem (Whig, Kasula, Jain, & Sharma, 2024). Furthermore, ML applications in data management are crucial in refining data processing algorithms and modeling techniques to ensure both compliance and efficiency (Pansara, Kasula, & Whig, 2024).

Overall, integrating machine learning into data retention and lifecycle management strategies is a promising avenue for creating more sustainable and intelligent data ecosystems. The potential of these technologies is increasingly being recognized, with applications ranging from predictive maintenance in supply chains to optimizing cloud-based data retention policies (Sugandha Agarwal et al., 2024; Alex Khang & Shashi Kant Gupta, 2024). By aligning data retention policies with advanced ML capabilities, organizations can not only reduce operational costs but also enhance decision-making processes, thus contributing to the long-term sustainability of their data management practices (Shashi Kant Gupta et al., 2023).