

A Governed, Rule-Based Preprocessing Framework with Decision Provenance for Auditable Data Analytics

Author: **Vivek Baburao Patil**

Under guidance of **Dr. Rama Bansode**

Abstract

While data preprocessing typically accounts for 50% to 80% of the effort in data science projects, it remains a largely ungoverned phase of the project lifecycle. Modern Automated Machine Learning (AutoML) tools prioritize predictive accuracy but often fail to provide the transparency and accountability required in regulated industries, leading to a "black-box" challenge. This study introduces a decision-centric five-layer framework designed to bridge this gap by combining YAML-based rule governance with granular decision provenance. By shifting from informal ad-hoc scripts to a formalized system of governed rules, the proposed architecture ensures that every data transformation is reproducible, justifiable and auditable. Experimental results across healthcare, finance, and climate science datasets indicate that the framework achieves total reproducibility and meets stringent regulatory requirements with minimal impact on computational performance.

Keywords: Data Preprocessing, Data Governance, Decision Provenance, Auditability, Design Science Research.

I. INTRODUCTION

The reliability of any analytical outcome depends on the quality of data preprocessing, which includes critical tasks such as normalization, data cleaning, and transformation. Despite being the most resource-intensive phase of the data lifecycle, preprocessing remains poorly standardized [1]. The emergence of Automated Machine Learning (AutoML) has introduced a notable "automation-governance paradox": while these systems streamline data preparation, they frequently obscure the underlying rationale for specific data modifications [2]. In high-stakes sectors such as finance and healthcare, such "black-box" approaches

are untenable because they lack the necessary audit trails and transparency in decision-making. Consequently, there is a pressing need for frameworks that offer not only automated efficiency but also robust, auditable governance and comprehensive provenance.

II. LITERATURE REVIEW

Contemporary research on data preparation generally follows two paths: automated optimization and rule-based quality control. Search-based tools, such as AlphaClean [4] and SAGA [8], excel in maximizing machine learning performance through automated pipeline generation; however, these methods often sacrifice interpretability for improved accuracy. Conversely, traditional rule-based systems primarily focus on operational data quality, such as schema validation, and rarely extend governance to analytical choices, such as outlier thresholds or imputation methods [5]. Although recent developments in provenance, such as the W3C PROV-DM model, provide standardized frameworks for tracking data lineage, they have not yet been effectively integrated into the statistical and logical decision-making processes inherent in preprocessing [10].

III. RESEARCH GAP

A significant limitation of current systems is the absence of "Decision Provenance" that is, the capability to articulate why a particular preprocessing strategy was selected over viable alternatives. Standard tools may record the occurrence of a transformation (lineage) but fail to document the underlying justification, specific rules triggered, or domain-specific constraints applied [1]. Furthermore, existing systems often lack the capacity to generate compliance-ready audit documentation required by frameworks such as HIPAA or GDPR, creating a gap between high-speed operational processing and regulatory accountability [9].

IV. METHODOLOGY

This study utilized the Design Science Research (DSR) methodology [7]. The framework was refined using a four-stage iterative process, as follows:

1. **Problem Identification:** Analyzing the existing literature and conducting interviews with practitioners.
2. **Artifact Design:** Development of a multilayered system architecture.
3. **Specification Development:** Creation of a YAML-based language for the formal definition of the rules.
4. **Empirical Evaluation:** Validation of the framework across diverse, cross-domain datasets. The objective was to develop a "utility-oriented" artifact that bridges the gap between the flexibility of manual coding and the rigorous structure of governed environments.

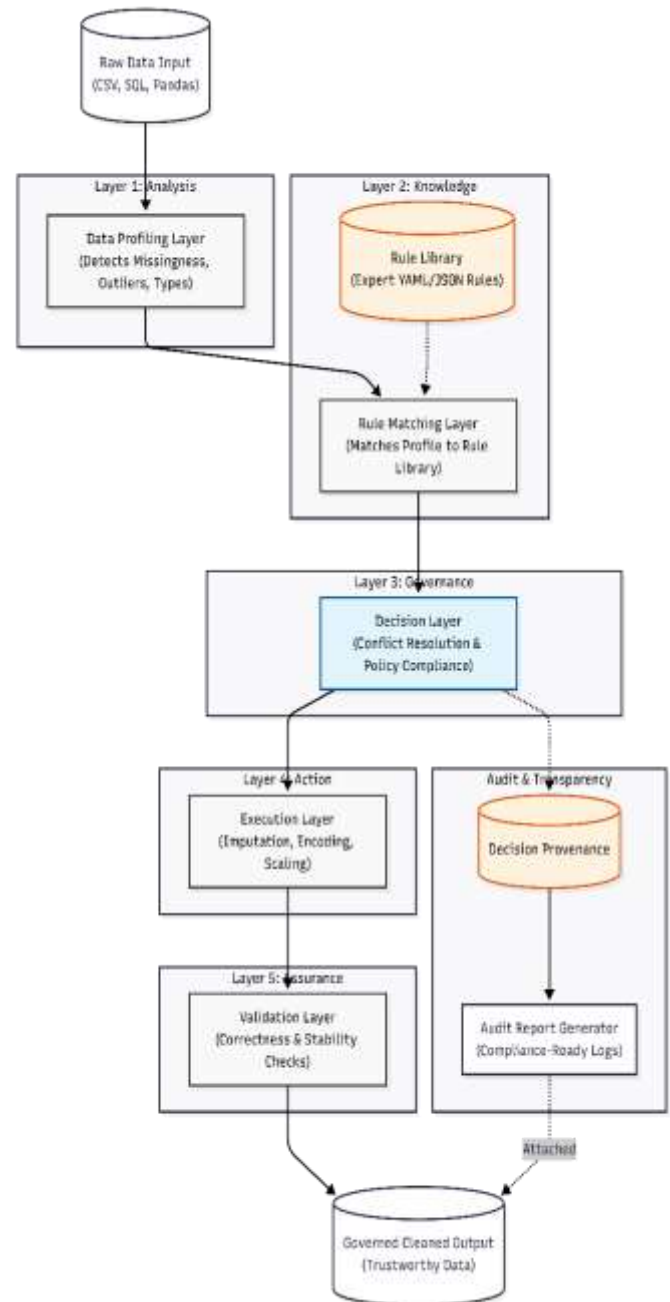
V. SYSTEM ARCHITECTURE

The proposed framework employs a five-layer decision-centric architecture to separate domain logic from the execution environment.

1. **Data Profiling Layer:** Performs a noninvasive analysis of data characteristics (e.g., skewness, missingness patterns, and data types) [2].
2. **Rule Matching Layer:** Connects the data profile with a curated library of domain-aware rules specified in YAML, identifying all potentially applicable transformations.
3. **Decision Layer:** Serving as the central logic engine, this layer handles conflict resolution (e.g., applying priority-based logic) and records the justification for every selected action [1].
4. **Execution Layer:** Implements the finalized transformations, such as Z-score scaling or KNN-based imputation, while maintaining comprehensive provenance logs.
5. **Validation Layer:** Conducts final quality assessments to ensure the statistical integrity and domain-level plausibility of the transformed data.

The architecture ensures that a clear path exists from the raw input (SQL, CSV, or Pandas) to a governed,

trustworthy output. This includes an integrated Audit Report Generator that produces compliance-ready logs detailing decision provenance.



VI. EVALUATION

The efficacy of the framework was tested using datasets from healthcare (MIMIC III), finance (Lending Club), and climate science. The evaluation focused on three primary metrics.

1. **Reproducibility:** The system achieved 100% consistency, yielding identical results across various environments using the same rule configurations.
2. **Auditability:** A panel of professional data auditors verified the completeness and utility of the decision logs generated.

3. **Performance:** Compared to standard, ungoverned Python scripts, the framework introduced an overhead of less than 12% for datasets as large as 10GB.

VII. CONCLUSION

The Governed Rule-Based Preprocessing Framework mitigates the risks associated with "black-box" data preparation by embedding decision provenance directly into the analytical pipeline. By standardizing and formalizing rule-based governance, this approach allows researchers and industry practitioners to perform data preparation that is both computationally efficient and legally defensible. Future research will explore the application of this framework to multimodal and real-time streaming data.

REFERENCES

- [1] [A. Chapman et al., "Capturing and querying fine-grained provenance of preprocessing pipelines," *VLDB*, 2020.](#)
- [2] [A. Chapman et al., "Supporting Better Insights of Data Science Pipelines with Fine-grained Provenance," *ACM TODS*, 2024.](#)
- [3] [R. A. Sola Martínez et al., "Data preprocessing workflow for exhaled breath analysis," *Sci. Rep.*, 2020.](#)
- [4] [S. Krishnan and E. Wu, "AlphaClean: Automatic Generation of Data Cleaning Pipelines," *arXiv*, 2019.](#)
- [5] [N. Parulian and B. Ludaescher, "Trust the Process: Analyzing Prospective Provenance for Data Cleaning," *WWW*, 2023.](#)
- [6] [A. Wagner, "FAIRly big: A framework for computationally reproducible processing," *Sci. Data*, 2022.](#)
- [7] [A. R. Hevner et al., "Design Science in Information Systems Research," *MIS Quarterly*, 2004.](#)
- [8] [S. Siddiqi et al., "SAGA: A Scalable Framework for Optimizing Data Cleaning Pipelines," 2023.](#)
- [9] [A. K. Ojha, "Reproducibility Under Preprocessing Variability," *Zenodo*, 2025.](#)
- [10] ["Health information systems: W3C PROV-DM instantiation," *Book Chapter*, 2022.](#)
- [11] [T. Dasu and T. Johnson, *Exploratory Data Mining and Data Cleaning*, Wiley, 2003.](#)
- [12] [H. V. Jagadish et al., "Big Data and Its Technical Challenges," *Commun. ACM*, 2014.](#)
- [13] [M. Stonebraker and I. Ilyas, "Data Curation: The Next Big Challenge," *IEEE Data Eng. Bull.*, 2018.](#)
- [14] [J. Wang et al., "Atlantic: A Framework for Automated Data Preprocessing," 2021.](#)
- [15] [C. Batini, F. and Scannapieco, M. \(2016\), *Data and Information Quality*, Springer.](#)