Cleansera: An Intelligent Desktop Application for Domain-Specific Data Cleaning

Shubham Deshmukh, Om Bhise, Shubham Rao, Vishal Daware

[1,2,3,4]*Student, Dept. of Information Technology, MET's Institute of Engineering, Nashik, Maharashtra, India*

---***---

Abstract – We know in today's world data professionals spend so much of their time, almost 50-80% on manual cleaning, this blocks the path to analysis. This paper presents Cleansera, an AI-powered desktop application made to speed up this process. Cleansera introduces a context-aware cleaning approach, this enables the system to apply industry-specific standards for domains like finance, supply chain, manufacturing. A Retrieval-Augmented Generation (RAG) framework is used in the system, allows users to interact with their data using natural language commands. To ensure data integrity, the system uses a dual-checkpoint quality assurance mechanism which tracks data loss and master fields. Application is built using Electron.js, functioning completely local with data privacy are the prioritizes. Ensuring high accuracy in AI command clarification are the key objectives. Cleansera automates complex data cleaning, allowing data professionals to focus on analysis and make accurate data-driven decisions.

***Keywords**: Data Cleaning, Context-Aware Systems, Retrieval-Augmented Generation (RAG), Artificial Intelligence (AI), Data Preparation Automation, Dual-Checkpoint Quality Assurance, Domain-Specific Data Standards, Electron.js Application, AI Systems, Natural Language Interface, Knowledge Retrieval, Data Integrity Verification, Industry-Aware Rule Application, Data Loss Detection, Cross-Platform Desktop Application*

1.INTRODUCTION

In today's data-centric world, information is a key asset for decision-making and organizational strategy. To understand customer behavior, optimize operations, and plan for future growth, many businesses, small or large, rely on high quality data. However, the raw data collected by organizations is rarely ready for analysis. The data cleaning and preparation are essential yet complex processes, due to different quality standards across different industries. For example, the rules for financial transaction are different from those required for marketing or ecommerce.

Traditional data cleaning tool often can't handle this complexity, as they rely on one-size-fits-all approach. Due to this approach the domain-specific requirements, compliance requirements, and standards that each industry sector demands are ignored. Because of this, data professionals have to set up the validation rules manually, a process that takes time and can easily lead to mistakes. These tools lack intelligence to understand the context of the data for example a "date" field in a financial ledger is treated identically to a "date" field in a supply chain manifest. This becomes even more difficult for professionals who handle data across different industries. Each domain whether it's Banking, Financial Services, and Insurance or supply chain management has their own rules and standards that needs to be followed during data preparation.

To overcome these inefficiencies, this study introduces Cleansera, an innovative desktop application that performs context-aware data cleaning. Cleansera is an AI-powered tool that understands and adapts to specific industry contexts by applying domain-specific data cleaning strategies. The system uses a Retrieval-Augmented Generation (RAG) framework, this allows the system to access external industry knowledge bases and best practice resources. This makes sure that all cleaning processes follow the right industry standards, moving past generic rules to create a smarter and more efficient data preparation solution.

2. LITERATURE REVIEW

The automation of data cleaning using large language models (LLMs) and artificial intelligence (AI) has been studied by several researchers in recent years. However, scalability, domain adaptation, transparency and real-time quality verification remain major obstacles for most current systems.

The effectiveness and scalability of traditional data cleaning and preprocessing tools were evaluated by Pedro Martins et al. In 2025. Their research showed that although these tools work well on structured datasets, they do not take domain-specific validation and contextual understanding into account. To get around this limitation, Cleansera offers context-aware intelligence that modifies intelligence that modifies cleaning processes based on supply chain, manufacturing and BFSI Domains.

AutoDCWorkflow, an LLM-based system that automatically creates data cleaning workflows, was presented by Lan Li et al. In 2025. The system needed human validation for most of its decision and lacked domain awareness, even though it decreased manual labor. To enable automation, domain-accurate cleaning without constant human supervision, Cleansera improves on this by integrating Retrieval-Augmented Generation (RAG) to access industry-specific knowledge bases.

The potential of LLMs as data preprocessors was examined by Elyas Meguellati et al. in 2025. Although they made some progress, their work lacked a systematic approach to identifying context, which led to poor results across industries. Cleansera, on the other hand, offers an automated context detection module that applies relevant industry-specific rules dynamically by identifying value distributions, column patterns, and data semantics.

Despite requiring a lot of processing power and falling to validate against domain-specific standards, Shuo Zhang et al. (2024) showed that LLMs are superior to conventional rule-based techniques in identifying data errors. This is addressed by Cleansera, which uses Electrons.js to implement a lightweight, effective Al model with local execution, guaranteeing performance optimisation and full offline functionality for settings that are sensitive to privacy.

RetClean, a retrieval-based framework created by Naeem et al. (2024), improved contextual accuracy by leveraging enterprise data lakes. However, it lacked generalisation for smaller or cross-domain datasets and was exclusively developed for large organisations. By providing a modular RAG architecture that scales across industries and facilitates local knowledge base integration, Cleansera expands this idea while preserving accuracy independent of enterprise infrastructure.

3. METHODOLOGY

Version controlled user interaction, industry-aware rule application, AI-driven context identification, RAG-based knowledge retrieval and dual-checkpoint quality assurance are all included in the methodology.

3.1 RAG Architecture Integration

In order to apply industry specific intelligence during data cleaning, Cleansera uses Retrieval-Augmented Generation. The RAG System obtains important information from:

- Industry standard repositories
- Validation rule libraries
- Templates for transformation
- Error pattern databases

As a result, the AI model can produce cleaning operations which are in line with domain-specific procedures. RAG improves the system's capacity to apply regulatory rules during cleaning based on the identified context, detect anomalies and standardise formats.

3.2 Advanced AI Chat Interface

The System allows users to communicate in plain English Through an AI-powered chat interface with version control. Without completely repeating workflows, users can duplicate, edit and rerun earlier prompts. The interface ensures continuous improvement of cleaning operations, supports selective editing and keeps track of all prompts and responses. By avoiding workflow resets, this conversation approach improves user experience and boosts productivity.

3.3 Dual-Checkpoint Quality Assurance System

First Checkpoint: Master Field Identification

AI is used by the system to determine which master fields are needed for data validation. Batch numbers in manufacturing, product codes in supply chain datasets, campaign IDs in marketing data and transaction IDs in BFSI are few examples. To make sure that cleaning procedures don't damage important identifiers, these Fields are highlighted for verification.

Second Checkpoint: Identification of Data Loss

The System Measures data loss by comparing the cleaned dataset with the raw dataset after the cleaning operations. This involves fields that have been updated, records that have been deleted, missing values and transformed entries. The System produces full report which explain the changes made and effects they have on the dataset. This ensures transparency and confidence in the cleaning procedure.

3.4 System Architecture

Electron.js is used in the development of Cleansera to enable cross-platform desktop deployment. The application includes:

Presentation Layer: Dashboards, chat systems and user interface.

Application Layer: Version control, RAG engine and context engine.

Data Layer: Vector databases, local storage and dataset management.

Complete local functionality is ensured by this architecture, which also shields private information from attackers.

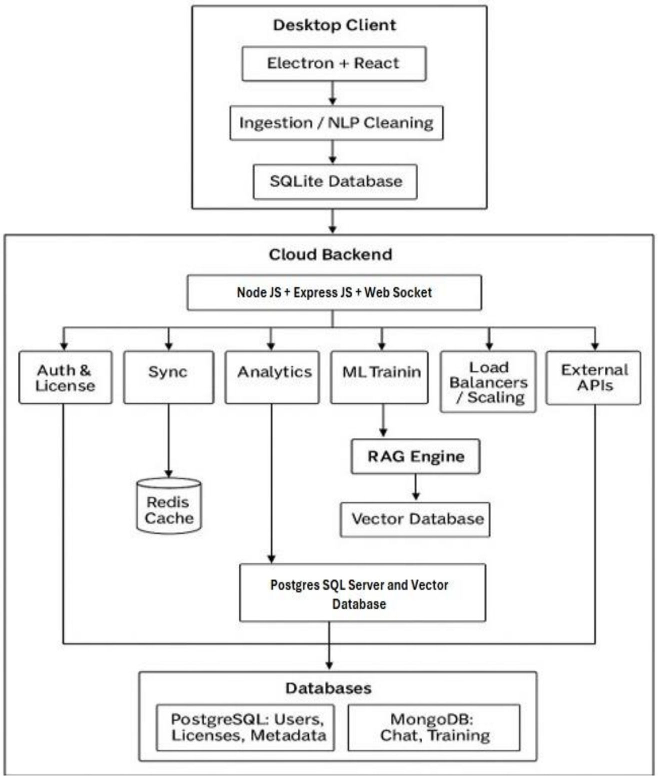

Fig -1: System Architecture

4. CONCLUSION

This paper presents Cleansera, a major advancement in data preparation that directly confronts the drawbacks of generic, one-size-fits-all cleaning tools. Cleansera is a next-generation system built on a Retrieval-Augmented Generation (RAG) architecture, establishing a signal, context-aware paradigm. The platform shows an ability to understand and implement domain standards of fields such as finance, supply chain, and manufacturing, thus ensuring that cleaning operations are not only automated but also operate according to the current best practices in the field. Central to the platform is a dual-checkpoint quality assurance system that verifies data integrity while transparently recording information loss. The main finding of this study is a platform that allows data professionals to greatly reduce cleaning time. As a result, we bring the researcher closer to turning raw data into meaningful insights.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Prof. Kanchan Dhomse and Prof. Kishor Mahale for invaluable guidance, constant encouragement, and insightful feedback throughout the development of this research work. Her expertise and support have played a crucial role in shaping this project and ensuring its successful completion.

We also extend our thanks to the Department of Information Technology, MET Institute of Engineering, for providing the necessary resources and a supportive research environment. We appreciate the cooperation of all faculty members and peers who contributed directly or indirectly to this study.

REFERENCES

P. Martins, F. Cardoso, P. V´az, J. Silva, and M. Abbasi, ”Performance and Scal ability of Data Cleaning and Preprocessing Tools: A Benchmark on Large Real World Datasets,” MDPI, 2025.

L. Li, L. Fang, B. Lud¨ascher, and V. I. Torvik, ”AutoDCWorkflow: LLM-based Data Cleaning Workflow Auto-Generation and Benchmark,” arXiv preprint, 2025.

E. Meguellati et al., ”Are LLMs Good Data Preprocessors?,” 2025.

S. Zhang, Z. Huang, and E. Wu, ”Data Cleaning Using Large Language Models,” arXiv preprint, 2024.

M. Naeem et al., ”RetClean: Retrieval-Based Data Cleaning Using LLMs and Data Lakes,” arXiv preprint, 2024.

L. Biester et al., ”LLMClean: Context-Aware Tabular Data Cleaning via LLM Generated OFDs,” 2024.

S. Zhang, Z. Huang, and E. Wu, ”Cocoon: Data Cleaning Using LLMs,” arXiv preprint, 2024.

W. Ni et al., ”IterClean: Iterative Data Cleaning with LLMs,” 2024.

F. Ahmadi, Y. Mandirali, and Z. Abedjan, ”Accelerating the Data Cleaning Sys tems Raha and Baran through Task and Data Parallelism,” VLDB Workshop, 2024.

J. Choi et al., ”Multi-News+: Cost-efficient Dataset Cleansing via LLM-based Data Annotation,” in Proceedings of EMNLP, 2024.

E. Rahm and H. H. Do, ”Data Cleaning: Problems and Current Approaches,” IEEE Bulletin of the Technical Committee on Data Engineering, vol. 23, no. 4, pp. 3-13, 2000.

P. Lewis et al., ”Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in Proceedings of NeurIPS, 2020.

Y. Gao et al., ”Retrieval-Augmented Generation for Large Language Models: A Survey,” arXiv preprint arXiv:2312.10997, 2023.

J. Devlin et al., ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proceedings of NAACL-HLT, 2019.

T. Brown et al., ”Language Models are Few-Shot Learners,” in Proceedings of NeurIPS, 2020.