



Enhancing Fault-Tolerant ETL Pipelines Through AI-Driven Predictive Maintenance: A Corporate Framework for Improved Data Quality and Integration

Wickramaarachchi W. A. P. C. K.
MS23017856

A THESIS
SUBMITTED TO
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

December 2024

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Prof. Samantha Thelijagoda

Approved for MSc. Research Project:



MSc. Programme Co-ordinator, SLIIT

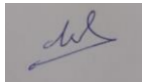
Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign:



Chathurindu Kaushalya

Date: 2024-11-09

ABSTRACT

Enhancing Fault-Tolerant ETL Pipelines Through AI-Driven Predictive Maintenance: A Corporate Framework for Improved Data Quality and Integration

Wickramaarachchi W.A.P.C.K.

MSc. in Information Technology

Supervisor: Prof. Samantha Thelijagoda

December 2024

Maintaining high data quality and consistency across various sources is essential for making informed and effective decisions in today's data-centric environment. This research presents an AI-driven approach to enhance fault tolerance within ETL (Extract, Transform, Load) pipelines, aiming to improve data quality through predictive maintenance mechanisms. The proposed ETL framework automates data cleaning, standardization, and error handling, utilizing machine learning and natural language processing (NLP) techniques to identify and resolve data inconsistencies in real time.

By integrating AI models into each phase of the ETL process, the pipeline demonstrates resilience against common data irregularities across varied formats, such as dates, numbers, and text. A unique feature of this approach is its predictive maintenance capability, where machine learning algorithms proactively address potential faults before they escalate, reducing downtime and increasing overall system reliability. Key components include LSTM-based models for date and text standardization, anomaly detection mechanisms for fault tolerance, and an automated error logging system to streamline data auditing processes. Results from experimental evaluations show that the AI-driven pipeline achieves significant improvements in data consistency and error detection, with up to a 98% reduction in inconsistencies for critical data fields. Despite some limitations, including resource intensity and sensitivity to rare data patterns, this research highlights the potential of AI-augmented ETL systems to meet the growing demand for robust data integration solutions in corporate environments. The findings suggest that AI-driven fault-tolerant ETL pipelines can play a pivotal role in advancing data quality management, enabling organizations to make data-driven decisions with greater confidence.

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to those who have supported and guided me throughout the course of this research.

Firstly, I am deeply indebted to my supervisor, whose invaluable insights, guidance, and encouragement have been instrumental in bringing this research to fruition. I am equally grateful to my lecturers, whose teachings and expertise have laid a strong foundation for this work, providing me with the skills and knowledge necessary to undertake this journey.

A special thank you goes to my wife, whose unwavering support, patience, and encouragement have been a constant source of strength. Her belief in my abilities has been a great motivation, especially during challenging times. I am also sincerely grateful to my parents and brother, whose continuous encouragement, love, and support have been essential throughout my academic journey.

I would also like to extend my appreciation to my colleagues, who have shared valuable feedback, provided support, and enriched this experience with collaborative discussions. Their camaraderie and insights have greatly contributed to this research.

Thank you all for your unwavering support and belief in me.

TABLE OF CONTENTS

DECLARATION	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENT.....	iv
TABLE OF CONTENTS.....	v
List of Figures	viii
List of Tables	ix
Chapter 1 Introduction	1
1.1 Background	1
1.2 Problem Statement.....	3
1.3 Research Objective	6
1.3.1 Automated Data Cleaning and Standardization	6
1.3.2 Automatic Mapping of CSV/Excel Headers to Database Table Headers	7
1.3.3 Dynamic Table Creation in the Data Warehouse.....	7
1.3.4 Ensuring Fault Tolerance and Operational Continuity.....	8
1.3.5 Minimizing Manual Intervention	8
1.4 Research Significance.....	9
1.4.1 Solving Key Problems in Traditional ETL Systems	10
1.4.2 Improving Data Quality	10
1.4.3 Building a Fault-Tolerant Data Pipeline	10
1.4.4 Adaptability and Scalability for Growing Data Needs	11
1.4.5 Reducing Manual Intervention and Costs.....	11
1.4.6 Enabling Real-Time Data Processing	11
1.4.7 Broader Impact on Data Engineering.....	12
1.5 Scope of the Study	12
1.5.1 Data Sources and Scope of Integration.....	12
1.5.2 Automation of Data Cleaning and Error Handling	13
1.5.3 Dynamic Table Creation and Schema Adaptation	13
1.5.4 Focus on Fault-Tolerance in ETL Pipelines	13
1.5.5 System Evaluation and Performance Metrics	14
1.5.6 Limitations of the Study	14
1.6 Thesis Structure	15
1.6.1 Chapter 1: Introduction	15
1.6.2 Chapter 2: Literature Review	15
1.6.3 Chapter 3: Methodology	16

1.6.4 Chapter 4: System Implementation	16
1.6.5 Chapter 5: Results and Evaluation	16
1.6.6 Chapter 6: Discussion	16
1.6.7 Chapter 7: Conclusion and Future Work.....	17
Chapter 2 Literature Review	18
2.1 ETL Processes: Traditional Approaches and Limitations.....	18
2.2 Fault Tolerance in ETL Pipelines.....	19
2.3 Automation in ETL Processes	20
2.4 AI-Driven Data Cleaning and Standardization.....	20
2.5 Dynamic Schema Matching and Table Creation	21
2.6 Gaps in Existing Research.....	22
Chapter 3 Methodology	26
3.1 Data Collection	26
3.1.1 Kaggle and GitHub Datasets.....	26
3.1.2 Generative AI (GenAI)	27
3.1.3 Ethical Considerations.....	27
3.2 Data Preprocessing	27
3.2.1 Data Cleaning	28
3.2.2 Data Splitting.....	28
3.2.3 Normalization.....	29
3.3 System Architecture Overview	29
3.3.1 Data Ingestion	29
3.3.2 Data Cleaning & Standardization	30
3.3.3 Schema Matching & Dynamic Table Creation.....	30
3.3.4 Data Warehouse	31
3.4 Model Development and Algorithms.....	31
3.4.1 Header Matching Model	31
3.4.2 Data Type Validation Model	32
3.4.3 Data Cleaning and Standardization Model	32
3.4.4 Dynamic Table Creation Model.....	36
3.4.5 Evaluation Metrics and Model Performance	38
Chapter 4 Results and Evaluation	41
4.1 Data Preprocessing Results.....	42
4.1.1 Dataset Overview	42
4.1.2 Preprocessing Outcomes	43
4.2 Model Training and Evaluation	45

4.2.1 Date Standardization Model	45
4.2.2 Text Cleaning Model	50
4.3 Integrated Pipeline Evaluation.....	54
4.3.1 Error Handling	54
4.3.2 Fault Tolerance.....	55
4.3.3 Processing Time and Efficiency	56
Chapter 5 Discussion.....	58
5.1 Achievements and Key Findings.....	58
5.1.1 Improvement in Data Consistency.....	58
5.1.2 Model Accuracy and Robustness	58
5.1.3 Fault Tolerance and Error Handling	59
5.2 Limitations.....	59
5.2.1 Model Dependency on Training Data	59
5.2.2 Ambiguity in Text and Date Interpretations	59
5.2.3 Resource and Computational Limitations.....	60
5.3 Comparison with Existing Solutions.....	60
5.4 Implications for Data Quality and Fault Tolerance in ETL.....	62
5.5 Future Work	63
5.5.1 Expanding Dataset Diversity	63
5.5.2 Advanced NLP Techniques for Ambiguous Data.....	63
5.5.3 Optimization for Scalability.....	63
Chapter 6 Conclusion and Future Work.....	64
References	66
Appendix	68
Appendix 1: Model Architecture and Hyperparameters	68

List of Figures

Figure 1: Training and validation accuracy for date standardization model	45
Figure 2: Training and validation loss for date standardization model	46

List of Tables

Table 1: Summary of the literature review.....	25
Table 2: Distribution of data formats in the training dataset.....	42
Table 3: Consistence records before and after.....	44
Table 4: Error analysis of date standardization model before fine-tuning.....	47
Table 5: Error analysis of date standardization model after fine-tuning.....	48
Table 6: Error analysis of text cleaning model.....	51
Table 7: Text cleaning model performance comparison before and after fine-tuning	53
Table 8: Error correction in the integrated pipeline	54
Table 9: Pipeline processing time and efficiency.....	56