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 Thelijjagoda

Approved for MSc. Research Project:

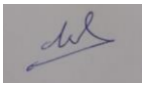
MSc. Programme Co-ordinator, SLIIT

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is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources).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 Thelijjagoda

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

List of Figures

List of Tables