VIETNAM NATIONAL UNIVERSITY OF HO CHI MINH CITY

INTERNATIONAL UNIVERSITY

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

# Anomaly Detection in HDFS Logs using Machine Learning Integrated with LLM-Based Mitigation

By

Nguyen Hoang Quan

The thesis submitted to School of Computer Science and Engineering in partial fulfillment of the requirements of the degree of Bachelor of Engineering of Information Technology

**Ho Chi Minh, Viet Nam**

**2025**

# Anomaly Detection in HDFS Logs using Machine Learning Integrated with LLM-Based Mitigation

APPROVED BY: ______________________________

# ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to the School of Computer Science and Engineering for providing a supportive academic environment and the essential resources that made this thesis possible.

I am deeply thankful to my advisor, Dr. Le Hai Duong, for their invaluable guidance, encouragement, and constructive feedback throughout the development of this thesis. Their expertise and commitment have greatly enriched both my technical understanding and research direction.

I would also like to extend my appreciation to all lecturers and staff members who have contributed to my academic journey with their instruction, support, and mentorship. Special thanks go to my family and friends for their constant support, motivation, and understanding during the preparation of this work. Lastly, I would like to acknowledge the developers and the researchers behind the resources for providing the foundation upon which this project was built.

This thesis would not have been possible without the guidance, resources, and support from all of the above.

# Contents

# List of Tables

# List of Figures

# ABSTRACT

Anomaly detection is essential for managing today's large-scale distributed systems, where system logs are a key resource for identifying unusual behavior. Traditionally, system operators relied on manual inspection methods such as keyword searches and rule-based matching. However, due to the massive volume and complexity of modern system logs, manual approaches are no longer practical. To tackle this, many automated log-based anomaly detection methods have been proposed. Still, developers often struggle to choose a suitable method, as there hasn't been a clear comparison of these approaches.

In this research, I will propose a solution to addresses the fundamental challenge of automated log analysis. Specially, the research tackles three interconnected problems: (1) automated parsing of diverse log formats into structured templates, (2) anomaly detection in high-volume log streams, and (3) generation of contextual, actionable recommendations for identified issues.

Previous research has established some foundational approaches including the algorithms for log parsing and machine learning techniques for anomaly detection. However, existing solutions typically focus on individual components rather than providing end-to-end integration. Most academic implementations lack production-ready deployment architectures, user-friendly interfaces, and the integration of modern Large Language Models (LLMs) for intelligent recommendations—creating a significant gap between theoretical algorithms and practical deployment. Therefore, this research is important since it close the gap between academic log analysis algorithms and production-ready systems.

Furthermore, the research establishes a framework for integrating emerging LLM capabilities into traditional system administration workflows, suggesting broader implications for AI-assisted DevOps practices. The findings indicate that intelligent automation of log analysis is not only technically feasible but can significantly enhance organizational capabilities in system reliability, security monitoring, and operational efficiency.

# Chapter 1

# Introduction

## 1.1 Background of the study

In modern software systems, log files serve as critical sources of information for system monitoring, debugging, troubleshooting, and security analysis. As applications or systems scale and increase its complexity, the volume of generated log data gets bigger exponentially. Therefore, traditional manual approaches for log analysis like manually examine through log files using basic tools such as grep or text editors, has become less efficient and more defective [1]. As the result, it is becoming more difficult to detect anomalies within large scale system.

Over the year, a lot of automated log-based methods have been introduced to help detecting system anomalies. These approaches usually require raw log preprocessing techniques, feature extraction and machine-learning-based algorithms for processing vast amounts of unstructured log data efficiently, identifying potential issues before they escalate into critical failures. Recent advancements in natural language processing and large language models (LLMs) have also increased the potential for intelligent log analysis systems. However, despite these development, traditional machine-learning techniques and LLM-based approaches have yet to be efficiently integrated into a single, cohesive log analysis framework.

This study focuses on leveraging existing machine learning approaches where log parsing and anomaly detected are used, and augmenting them with large language model to provide actionable insights and helpful recommendations.

## 1.2 Problem Statement

Despite the important role of log analysis that play in making the system more secure and reliable, several significant challenges still persist in current practices:

- *Limited interpretability.* In log anomaly detection, the ability to interpret model's outputs is important for people who work as system administrators or analysts to effectively action to the alerts. They need to understand which log entries may be responsible for the detected abnormality. Yet, many traditional approaches only provide basic classified prediction without any explanation. As a result, engineers still have to perform further manual root cause analysis which in large-scale and complex systems becomes an very time-consuming and heavy task.

- *Poor adaptability.* Many existing methods rely on a predefined set of log event templates during feature extraction phase (This phase will use the set of log event templates generated by log parsing to create numerical features for machine learning models [2]). However, as applications scale up and expand in term of feature, new and unseen log events will definitely appear. Therefore, adapting to these changes require retraining models from scratch which make the systems become less practical in dynamic environments.

- *Poor adaptability.*

## 1.3 Objectives of the Study

This study will perform the investigation on existing anomaly-detection methodologies, compares the performance of different machine-learning models, and develops a practical framework that integrates large language models (LLMs) to automate and improve anomaly detection in real-world scenario.

## 1.4 Limitations of the Study

# Bibliography

[1] John Doe and Jane Smith. Anomaly detection in distributed systems. *Journal of Computing*, 10:1–15, 2023.

[2] Mohanad Sarhan, Siamak Layeghy, Nour Moustafa, Marcus Gallagher, and Marius Portmann. Feature extraction for machine learning-based intrusion detection in iot networks. *Digital Communications and Networks*, 10(1):205–216, 2024.