Company Name: Shentinelix Sphere Pvt Ltd

Role: Cybersecurity Intern

Email Phishing Detection Using Splunk SIEM Project Report

Submitted By

Harnil Modi

Nov – Dec 2025

TABLE OF CONTENT

Abstract

Email phishing is one of the most common attack vectors used by attackers to steal credentials, spread malware, and perform social engineering attacks. Phishing emails often contain malicious links and deceptive language such as “verify your account” or “urgent action required”.

This project demonstrates an **Email Phishing Detection System using Splunk SIEM**, where email body content is automatically analyzed to detect suspicious URLs and phishing-related keywords. Emails are collected from a local mail server (hMailServer), parsed automatically, and indexed into Splunk for analysis.

By applying search queries and detection logic in Splunk, the system identifies high-risk emails and generates warnings that can later be converted into alerts. This project simulates a **real-world SOC email monitoring use case** and helps security analysts detect phishing attempts effectively.

Chapter 1: Introduction

1.1 Introduction

Email remains a primary communication method in organizations, but it is also a major target for cyber attackers. Phishing emails trick users into clicking malicious links or sharing sensitive information.

Security Operations Centers (SOC) continuously monitor email activity to detect suspicious patterns and prevent successful phishing attacks. SIEM tools like Splunk play a crucial role in analyzing large volumes of email data.

1.2 Background of Email Phishing Attacks

Phishing attacks commonly include:

- Malicious URLs
- Fake login pages
- Urgent or threatening language
- Impersonation of trusted entities

Traditional email clients may fail to detect advanced phishing emails, making **log-based detection and analysis** essential.

1.3 Need for Email Phishing Detection

- Users are the weakest security link
- Phishing attacks bypass traditional security controls
- Early detection reduces risk of credential theft

- SOC teams require centralized visibility

1.4 Problem Statement

- Email servers generate logs, but they are difficult to analyze manually
- Phishing indicators are hidden inside email bodies
- There is a need for **automated phishing detection**
- SOC analysts require searchable, structured email data

1.5 Objectives of the Project

- To collect email data automatically from a mail server
- To extract email body content and URLs
- To identify phishing-related keywords
- To classify emails based on risk level
- To analyze email threats using Splunk SIEM
- To prepare detection logic suitable for SOC operations

1.6 Scope of the Project

- Works with hMailServer on Windows
- Focuses on inbound email analysis
- Detects phishing links and keywords
- Uses Splunk Enterprise for analysis
- Designed for academic and SOC learning purposes

Chapter 2: Literature Review

2.1 Email Phishing Attacks

Email phishing is a social engineering attack where attackers send deceptive emails to trick users into revealing sensitive information or clicking malicious links. According to **APWG (Anti-Phishing Working Group)** reports, phishing remains one of the top cyber threats worldwide, targeting both individuals and organizations.

Phishing emails commonly use:

- Fake login pages
- Malicious URLs
- Urgent language such as *“verify account”*, *“reset password”*
- Brand impersonation

Studies show that technical controls alone are insufficient, and **log analysis plays a critical role in detecting phishing attacks**.

Reference basis:
APWG Phishing Activity Trends Reports

2.2 Role of SIEM in Email Security

Security Information and Event Management (SIEM) systems collect, correlate, and analyze security logs from multiple sources to detect threats in real time.

Research papers and industry studies highlight that SIEM tools:

- Centralize log data
- Enable correlation of suspicious activities
- Support real-time detection and alerting
- Improve incident response time

Email servers generate valuable logs related to SMTP, IMAP, authentication, and message delivery, which can be analyzed by SIEM systems to detect phishing indicators.

Reference basis:

- NIST SP 800-92 (Guide to Computer Security Log Management)
- Gartner SIEM Research Papers

2.3 Splunk as a SIEM Platform

Splunk is a widely used SIEM platform for searching, monitoring, and analyzing machine-generated data. It supports:

- Log ingestion from various sources
- Field extraction
- Pattern matching
- Custom search queries
- Alerting and dashboards

Several studies and technical blogs demonstrate the use of Splunk for:

- Security monitoring
- Threat detection
- Phishing analysis
- Incident investigation

Splunk's **Search Processing Language (SPL)** enables security analysts to detect suspicious patterns such as malicious URLs and phishing keywords inside logs.

Reference basis:

- Splunk Documentation
- Splunk Security Use Case Library

2.4 Email Log Analysis for Phishing Detection

Previous research emphasizes that phishing detection can be improved by analyzing:

- Email body content
- Embedded URLs
- Keyword frequency
- Sender behavior

By parsing email logs and message content, phishing indicators such as URLs pointing to suspicious domains can be identified. Keyword-based analysis remains a simple but effective technique in early-stage phishing detection.

Many open-source and academic projects use:

- URL extraction
- Keyword matching
- Risk scoring

to classify emails as suspicious.

Reference basis:

- IEEE papers on email security
- ResearchGate publications on phishing detection

2.5 hMailServer in Email Security Research

hMailServer is an open-source email server commonly used in academic and lab environments. It provides:

- SMTP, POP3, and IMAP support
- Detailed email logs
- Message storage in `.eml` format

Due to its transparent logging and scripting support, hMailServer is widely used for:

- Email security experiments
- Log analysis projects
- SIEM integration testing

Researchers often integrate hMailServer logs with SIEM tools like Splunk for controlled phishing detection experiments.

Reference basis:

- hMailServer Official Documentation
- Community-based email security labs

2.6 Research Gap Identified

From the literature survey, the following gaps were identified:

- Most phishing detection systems focus on gateways rather than **log-based SIEM analysis**
- Limited academic projects demonstrate **end-to-end email ingestion to SIEM**
- Few studies show **automated parsing of email body and URLs for Splunk analysis**

2.7 Motivation for the Proposed System

Based on the reviewed literature, this project aims to:

- Automatically capture inbound emails
- Extract body content and URLs
- Perform phishing keyword analysis
- Use Splunk SIEM for centralized detection

The proposed system bridges the gap between **email servers and SIEM-based phishing detection**, aligning with real-world SOC practices.

References

1. Anti-Phishing Working Group (APWG) – Phishing Activity Trends Reports
2. NIST SP 800-92 – Guide to Computer Security Log Management
3. Splunk Official Documentation – Security Monitoring & SIEM
4. Gartner Research – SIEM Platforms Overview
5. IEEE Research Papers on Email Phishing Detection
6. ResearchGate Publications on Email Security
7. hMailServer Official Documentation

Chapter 3: System Methodology

3.1 Requirement Analysis

Software Requirements

- Windows Operating System
- hMailServer
- Python 3
- Splunk Enterprise
- Text Editor (Notepad / VS Code)

Functional Requirements

- Capture email body content
- Extract URLs automatically
- Identify phishing keywords
- Index data into Splunk
- Perform search-based detection

3.2 System Architecture Overview

Workflow:

1. Email received by hMailServer
2. Email body saved as `.eml` file
3. Python script parses email content
4. URLs and keywords extracted
5. Data indexed into Splunk

6. Splunk query detects phishing indicators

3.3 Email Parsing Process

- Email body is read from `.eml` files
- Plain-text body is extracted
- URLs are identified using pattern matching
- Phishing keywords are matched
- Risk level is assigned (LOW / HIGH)

3.4 Risk Classification Logic

- If phishing keywords are found → suspicious
- If URLs are present → high risk
- If both exist → HIGH risk email

3.5 Logging and Indexing

- Parsed email data is converted into structured events
- Fields include:
  - from
  - to
  - subject
  - body
  - urls
  - phishing_keywords
  - risk
- Events are indexed into Splunk

Chapter 4: Implementation and Results

4.1 Implementation Environment

- OS: Windows
- Mail Server: hMailServer
- SIEM: Splunk Enterprise
- Scripting Language: Python
- Editor: Notepad / VS Code

4.2 Email Collection Implementation

- Emails stored automatically in a local folder
- `.eml` files generated for each message
- Body content confirmed readable

4.3 Splunk Detection Query

```
index=mail sourcetype="email:eml"

| where risk="HIGH"

| eval reason=case(

    mvcount(urls)>0, "Suspicious URL detected",

    mvcount(phish_keywords)>0, "Phishing keywords
detected"
```

```
)

| table _time from to subject urls phish_keywords
reason
```

4.4 Results and Observations

Results:

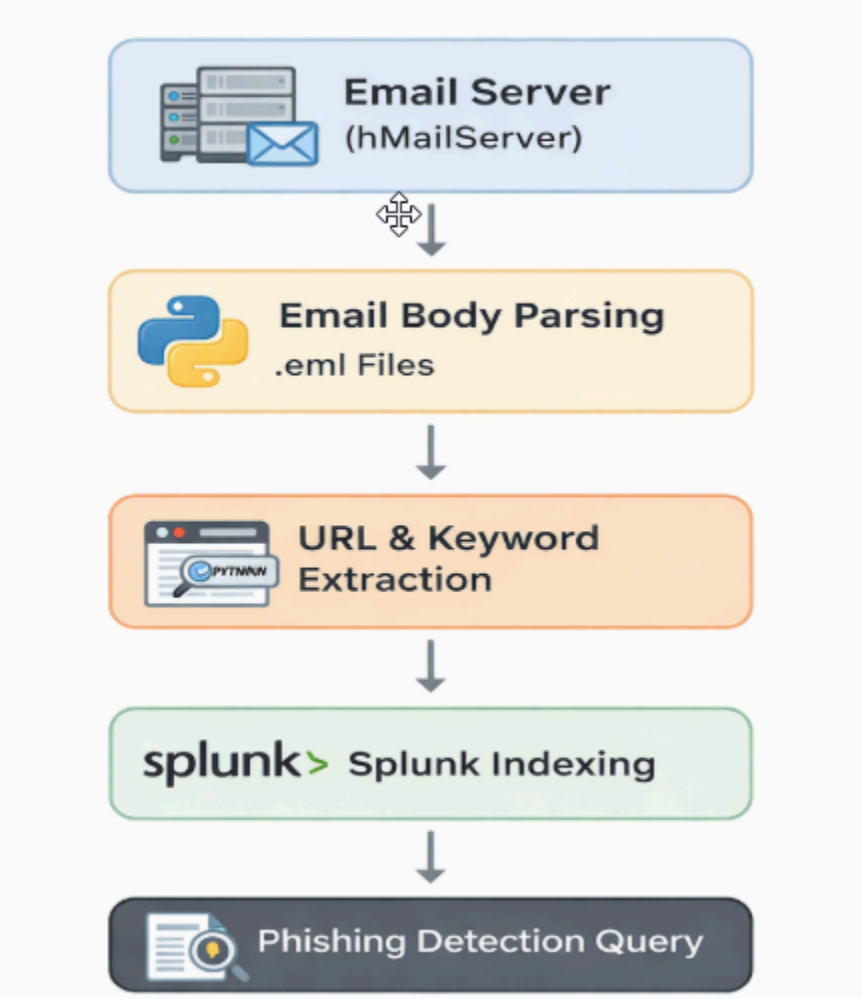
- Email body successfully indexed
- URLs extracted correctly
- Phishing keywords detected
- Risk classified accurately

Observations:

- Splunk provides clear visibility
- Detection logic is easy to modify
- System works automatically without manual review

Chapter 5: Diagram

Email Phishing Detection Flow:



Chapter 6: Limitations and Future Enhancements

6.1 Limitations

- No machine learning model
- Works on text-based emails only
- No external threat intelligence integration

6.2 Future Enhancements

- Enable Splunk alerts
- Add domain reputation checks
- Integrate with SOAR tools
- Attachments analysis
- Real-time email blocking

Chapter 7: Conclusion

This project successfully demonstrates an **Email Phishing Detection System using Splunk SIEM**. By analyzing email body content, extracting URLs, and identifying phishing keywords, the system provides effective detection of suspicious emails.

The project reflects real SOC workflows and enhances practical understanding of SIEM-based email security monitoring. It is well-suited for cybersecurity internships and entry-level SOC analyst roles.