PDFGUARD SENTINEL

PDF Malware Analysis & Detection Framework

Domain: Cybersecurity / Blue Team
Type: Practical Security Project
Submitted by: *Chirayu Paliwal*
Duration: Internship Practical Assignment
Tools & Technologies: Python, PDF Parsing Libraries, Static Malware Analysis, JSON Logging

1. INTRODUCTION

Modern malware delivery does not rely solely on executables. Documents—especially PDF files—have become a preferred attack vector due to their widespread trust, cross-platform compatibility, and ability to embed active content. Attackers routinely weaponize PDFs using embedded JavaScript, malicious actions, encoded payloads, and hidden objects to exploit user behavior and software vulnerabilities.

PDF-based malware is commonly used in phishing campaigns, invoice fraud, resume-based attacks, and targeted intrusions against enterprise users. These threats often bypass basic file-type filtering and rely on user interaction rather than direct exploitation.

PDFGUARD SENTINEL is a static PDF malware analysis framework designed to inspect PDF files at a structural level, identify suspicious and malicious elements, extract indicators of compromise, and generate security-relevant logs for analyst review.

2. OBJECTIVE OF THE PROJECT

• To analyze PDF files for malicious and suspicious structural elements
• To detect embedded JavaScript, launch actions, and risky PDF objects
• To extract indicators of compromise such as hashes, URLs, and suspicious strings
• To classify findings based on risk and behavior
• To generate structured, SOC-ready logs for security monitoring and investigation

2.1. PROBLEM STATEMENT

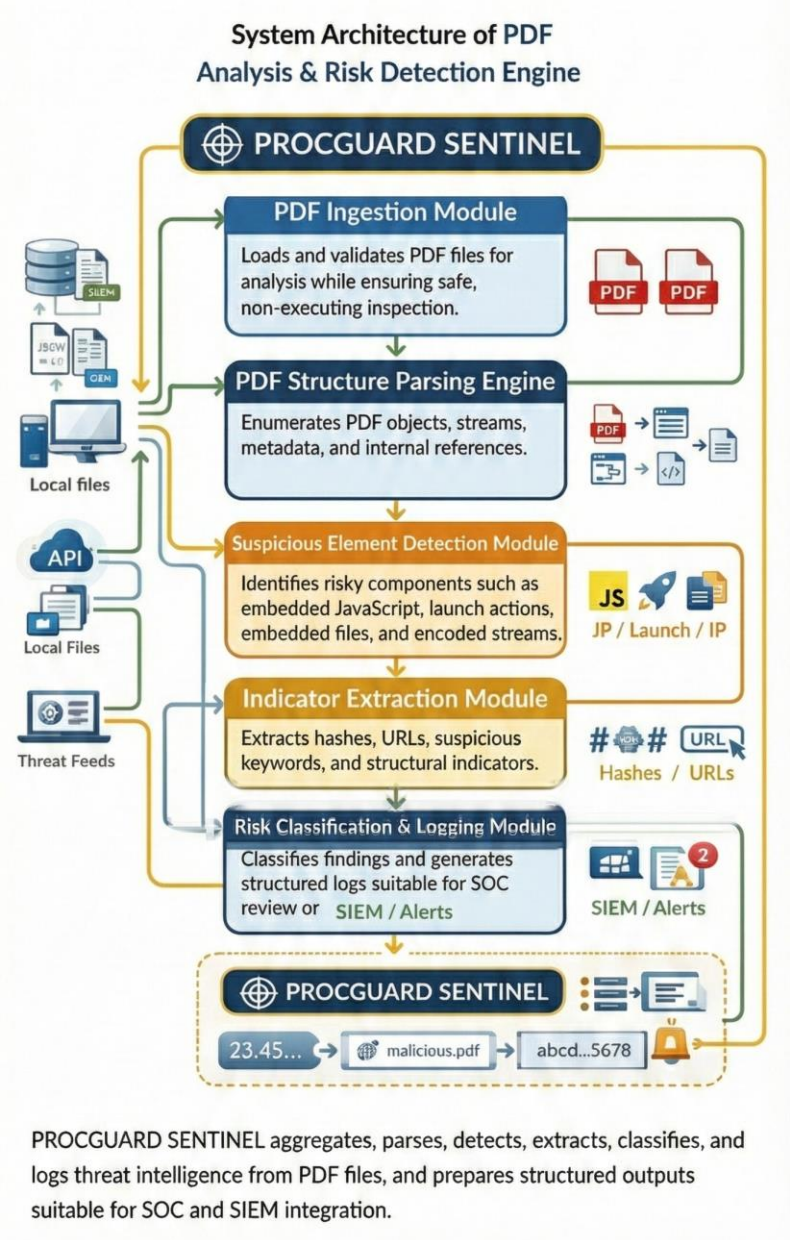
PDF files are often assumed to be safe due to their non-executable nature. In reality, the PDF format supports scripting, embedded files, external actions, and obfuscation techniques that attackers actively abuse.

Challenges with PDF malware analysis include:

1. Malicious logic hidden inside complex PDF object structures
2. Obfuscated JavaScript embedded within streams
3. Lack of visibility into document behavior without proper parsing
4. Overreliance on antivirus signatures rather than structural analysis

Many basic tools rely only on hash reputation or superficial scanning, which fails against novel or heavily obfuscated threats. This project addresses the gap by focusing on **structural inspection, behavior indicators, and forensic visibility**, rather than execution or signature-based detection.

3. ARCHITECTURE



PDFGUARD SENTINEL follows a modular static analysis architecture inspired by document inspection engines used in enterprise email security solutions and SOC pipelines.

3.1 ARCHITECTURE COMPONENTS

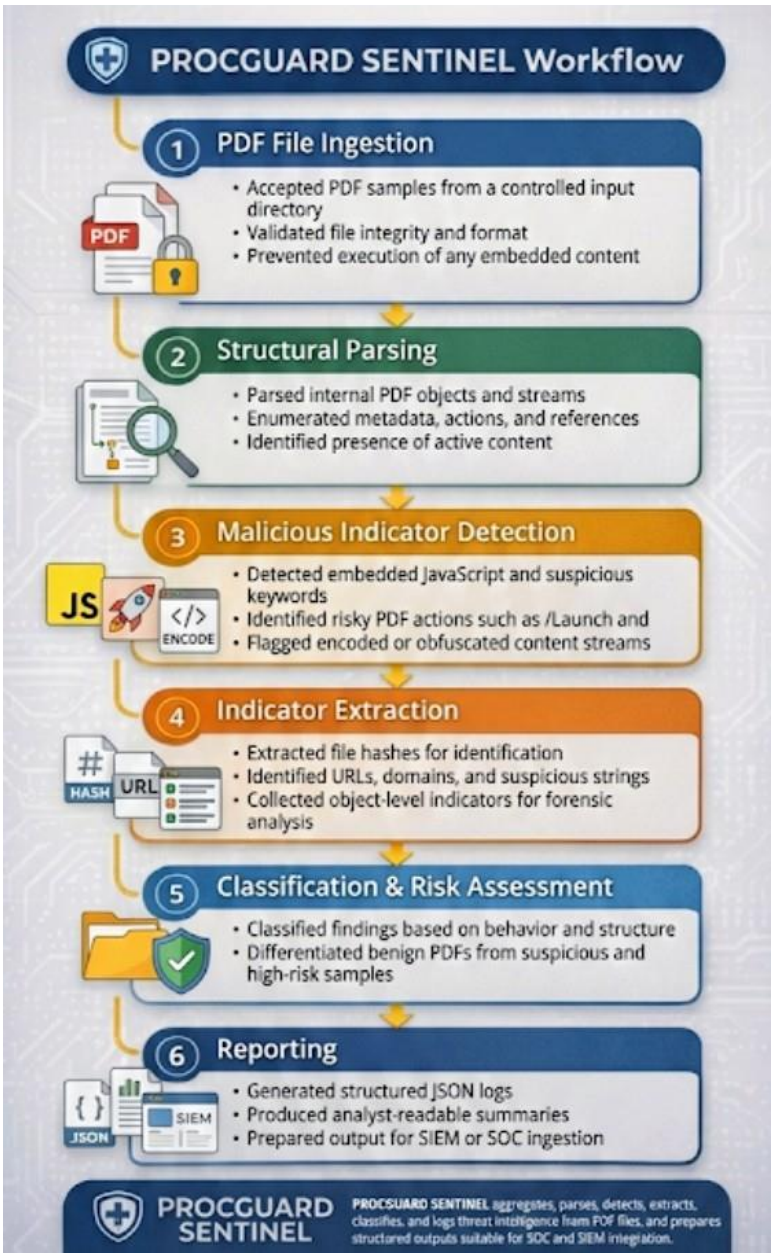
- **PDF Ingestion Module**
  Loads and validates PDF files for analysis while ensuring safe, non-executing inspection.

- **PDF Structure Parsing Engine**
  Enumerates PDF objects, streams, metadata, and internal references.

- **Suspicious Element Detection Module**
  Identifies risky components such as embedded JavaScript, launch actions, embedded files, and encoded streams.

- **Indicator Extraction Module**
  Extracts hashes, URLs, suspicious keywords, and structural indicators.

- **Risk Classification & Logging Module**
  Classifies findings and generates structured logs suitable for SOC review or SIEM ingestion.


PROCGUARD SENTINEL Workflow
1
PDF File Ingestion
PDF
• Accepted PDF samples from a controlled input directory
• Validated file integrity and format
• Prevented execution of any embedded content
2
Structural Parsing
• Parsed internal PDF objects and streams
• Enumerated metadata, actions, and references
• Identified presence of active content
3
Malicious Indicator Detection
JS
ENCODE
• Detected embedded JavaScript and suspicious keywords
• Identified risky PDF actions such as /Launch and
• Flagged encoded or obfuscated content streams
4
Indicator Extraction
HASH
URL
• Extracted file hashes for identification
• Identified URLs, domains, and suspicious strings
• Collected object-level indicators for forensic analysis
5
Classification & Risk Assessment
• Classified findings based on behavior and structure
• Differentiated benign PDFs from suspicious and high-risk samples
6
Reporting
JSON
SIEM
• Generated structured JSON logs
• Produced analyst-readable summaries
• Prepared output for SIEM or SOC ingestion
PROCGUARD SENTINEL
PROCSUARD SENTINEL aggregates, parses, detects, extracts, classifies, and logs threat intelligence from PDF files, and prepares structured outputs suitable for SOC and SIEM integration.

4. METHODOLOGY

Step 1: PDF File Ingestion

- Accepted PDF samples from a controlled input directory
- Validated file integrity and format
- Prevented execution of any embedded content

Step 2: Structural Parsing

- Parsed internal PDF objects and streams
- Enumerated metadata, actions, and references
- Identified presence of active content

Step 3: Malicious Indicator Detection

- Detected embedded JavaScript and suspicious keywords
- Identified risky PDF actions such as `/Launch` and `/OpenAction`
- Flagged encoded or obfuscated content streams

Step 4: Indicator Extraction

- Extracted file hashes for identification
- Identified URLs, domains, and suspicious strings
- Collected object-level indicators for forensic analysis

Step 5: Classification & Risk Assessment

- Classified findings based on behavior and structure
- Differentiated benign PDFs from suspicious and high-risk samples

Step 6: Reporting

- Generated structured JSON logs
- Produced analyst-readable summaries
- Prepared output for SIEM or SOC ingestion

5. TOOLS & TECHNOLOGIES USED

Tool / Technology	Purpose
Python 3.14	Core aggregation and processing logic
PDF Parsing Libraries	Structural inspection of PDF files
JSON	Structured logging and output
Visual Studio Code & POWERSHELL	Development environment Testing and validation

6. Threat Model and Assumptions

Attacker Assumptions

- PDFs are used as initial access vectors
- Malicious logic is hidden within document structure
- Obfuscation is used to evade basic scanning

Defender Assumptions

- Static analysis only (no execution or sandboxing)
- Focus on detection, visibility, and investigation
- Output intended for SOC analysts and monitoring systems

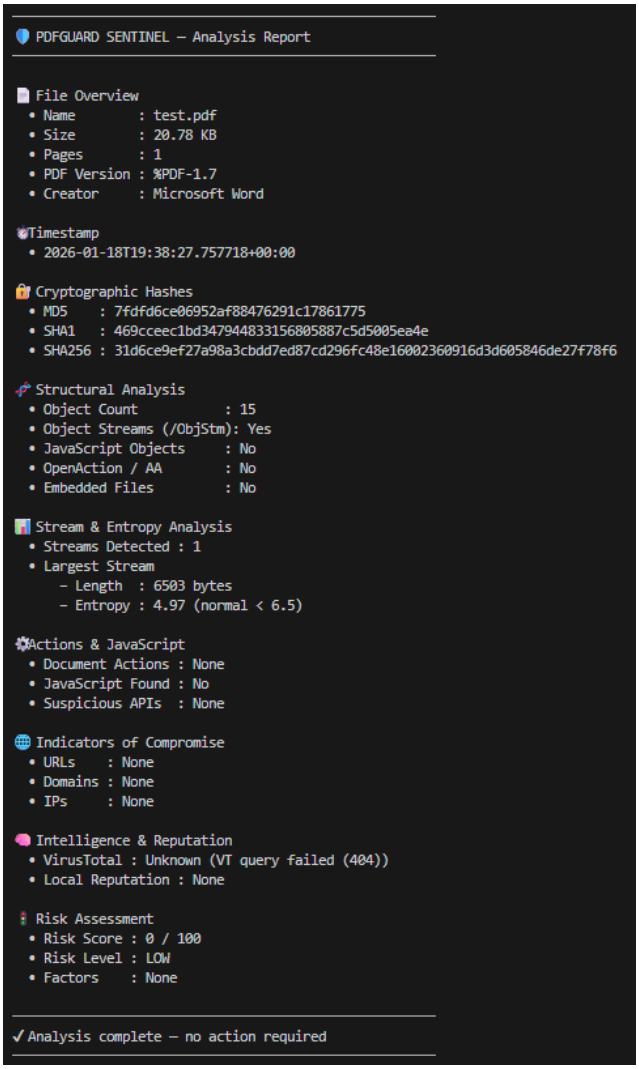
7. DETECTION & CORRELATION PHILOSOPHY

PDFGUARD SENTINEL does not treat PDF files as inherently safe. Instead:

- Structural behavior is prioritized over file appearance
- Presence of active content increases risk scoring
- Explainability and forensic visibility are emphasized
- Detection focuses on **what the document can do**, not what it claims to be

This mirrors real-world defensive document analysis used in SOCs and secure email gateways.

8. Observations



**Here VirusTotal query failed beacuse it has never seen this type of pdf.*

During testing, PDFGUARD SENTINEL successfully demonstrated:

- Detection of embedded JavaScript within PDF files
- Identification of suspicious actions and encoded streams
- Extraction of security-relevant indicators
- Clear differentiation between benign and suspicious documents

9. RESULTS

The system:

- Analyzed PDF files without executing embedded content
- Identified malicious and suspicious structural elements
- Extracted indicators of compromise for investigation
- Generated structured, analyst-ready output
- Simulated a real-world PDF malware inspection workflow

10. CONCLUSION

Through the development of PDFGUARD SENTINEL, I gained a practical understanding of how document-based malware is analyzed in real-world security environments.

This project demonstrated that PDF malware detection is not about blindly trusting file extensions—it is about understanding document structure, identifying risky behavior, and extracting actionable intelligence. By focusing on static analysis and explainable detection, PDFGUARD SENTINEL mirrors how SOC teams and email security solutions assess document threats before they reach end users.

The project strengthened my blue team skillset by combining malware analysis, Python development, structured logging, and defensive security thinking. PDFGUARD SENTINEL represents a realistic and practical approach to document-based threat detection in modern security operations.