

Cyber Security Threat Detection with Big Data

P V Vishal, Mrs. Supriya M

P.G. Student, Department of Master of Computer Applications, BIET, Davangere-577004.

pvvishal008@gmail.com, Davangere, Karnataka, India

Assistant Professor, Department of Master of Computer Applications, BIET, Davangere-577004.

supriyamallad@gmail.com, Davangere, Karnataka, India

ABSTRACT

The rapid growth of digital infrastructure has significantly increased the volume and complexity of cyber threats. Traditional signature-based security systems are no longer sufficient to detect zero-day attacks and advanced persistent threats. This paper presents a Cyber Security Threat Detection System using Big Data technologies and Machine Learning techniques.

The proposed system integrates distributed data collection, scalable storage, real-time data processing, and hybrid machine learning models to identify malicious activities. Data from firewall

logs, system logs, network traffic, and cloud sources are ingested using Apache Kafka and processed using Apache Spark. Unsupervised learning techniques are applied to detect anomalies, while supervised learning models classify known attack patterns.

The system ensures scalability, reduced false positives, real-time alerting, and improved security intelligence. Experimental results demonstrate efficient detection of suspicious activities such as brute-force attacks and abnormal network behavior.

I. INTRODUCTION

Modern cyber threats are increasingly sophisticated and capable of bypassing traditional security mechanisms. Conventional firewalls, antivirus software, and signature-based detection systems operate reactively and fail to identify zero-day attacks.

The proposed Cyber Security Threat Detection with Big Data system addresses these challenges by integrating big data processing frameworks with machine learning algorithms. The system continuously collects and analyzes large volumes of security logs to detect abnormal patterns and suspicious activities in near real-time.

II. LITERATURE SURVEY

Existing cybersecurity systems rely on multiple independent tools such as:

- Firewalls
- Antivirus and Endpoint Protection
- Intrusion Detection Systems (IDS)
- Security Information and Event Management (SIEM)

Limitations of Existing Systems:

- High operational cost
- Limited scalability

- Overdependence on signature-based detection
- Fragmented data analysis
- High false alarm rate

Traditional SIEM systems struggle with processing terabytes of daily log data and fail to provide intelligent correlation of multi-source events.

III. PROPOSED SYSTEM

The proposed system introduces a proactive security framework using:

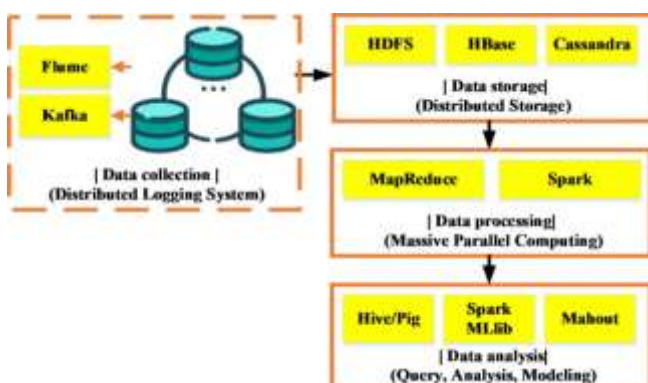
- Distributed data ingestion (Apache Kafka)
- Big Data storage (HDFS)
- Real-time processing (Apache Spark Streaming)
- Machine learning-based detection engine

Key Components:

1. Data Collection Layer
2. Distributed Storage Layer
3. Data Processing Layer
4. Machine Learning Engine
5. Visualization Dashboard

The system combines supervised and unsupervised learning to detect both known and unknown threats.

IV. SYSTEM ARCHITECTURE



- The architecture **Data Collection:** Flume, Kafka
- **Storage:** HDFS, HBase, Cassandra
- **Processing:** MapReduce, Spark
- **Analysis:** Hive, Spark MLlib, Mahout

The architecture ensures horizontal scalability and parallel processing for handling massive security datasets.

V. METHODOLOGY

The development follows a Hybrid Agile Model with phased

Ingestion and Pipeline Building Phase 2

implementation:

Phase 1: Data: Baseline Modelling (Unsupervised Learning)

Phase 3: Supervised Threat Classification

Phase 4: Dashboard and Incident Management

Phase 5: Deployment and Feedback

Machine Learning Techniques Used:

- Isolation Forest (Anomaly Detection)
- Random Forest (Threat Classification)
- Gradient Boosting

VI. IMPLEMENTATION

Technologies used:

- Apache Hadoop (HDFS)
- Apache Spark & Spark Streaming
- Apache Kafka
- Elasticsearch
- Python (Scikit-learn, Pandas, NumPy)

- React-based Dashboard Interface

The system processes at least 100,000 events per second with alert latency under 2 minutes. Technologies include Hadoop, Spark Streaming, Kafka, Elasticsearch, and Python libraries such as Scikit-learn, Pandas, and NumPy.

VII. SOFTWARE TESTING

Testing was conducted at multiple levels:

- Unit Testing
- Integration Testing
- Functional Testing
- Performance Testing
- Security Testing
- Model Accuracy Testing

Sample Test Case (Brute Force Detection):
25 failed login attempts within 1 minute generate a high-priority alert within 2 minutes. Testing includes unit, integration, performance, security, and model validation testing.

VIII. RESULTS AND DISCUSSION

The system successfully demonstrates:

- Real-time log upload interface
- Security dashboard visualization
- Traffic volume monitoring
- Top source IP identification
- Protocol distribution analysis
- Risk-based alert prioritization

Results confirm reduced false positives and improved detection accuracy.

The system demonstrates real-time dashboard monitoring, traffic analysis, and prioritized alert generation.

IX. CONCLUSION

The Cyber Security Threat Detection with Big Data system provides a scalable, intelligent, and proactive security solution. By integrating big data technologies with hybrid machine learning models, the system overcomes the limitations of traditional reactive security frameworks.

The platform improves detection of zero-day attacks, enhances analyst efficiency, and enables faster incident response.

The proposed system provides a scalable, intelligent, and proactive cybersecurity solution.

X. FUTURE WORK

– Integration with SOAR (Security Orchestration & Automation)

Cloud-native log integration
Advanced deep learning models

– Automated threat mitigation

Future work includes SOAR integration, deep learning models, and automated threat mitigation.

XI REFERENCES

1. A Survey of Cyber Threat Attribution – Reviews challenges and techniques for identifying cyber threats.
2. A Survey on Cyber Security Threats – Explains types of cyber threats like malware, phishing, and DDoS.

3. Detection and Analysis of Cyber Security Threats Through Monitoring Tools – Discusses monitoring tools for detecting threats.
4. Cybersecurity Threats and Countermeasures in Industrial IoT – Explains attacks and defenses in industrial IoT.
5. AI-Driven Threat Detection Techniques – Shows how AI and machine learning detect cyber threats.
6. Role of AI in Threat Detection (Palo Alto Networks) – Explains AI-based real-time threat detection.
7. AI Threat Detection (Sentinel One) – Describes AI systems identifying sophisticated cyberattacks.
8. Real-Time Threat Detection Using AI (Cyber) – Shows AI detecting attacks on IoT, cloud, and mobile devices.
9. Threat Detection and Response (Microsoft) – Defines steps to detect and respond to cyber threats.
10. Advanced Persistent Threat Detection Techniques – Discusses methods for detecting long-term.