

PhD, Associate Professor Boiarinova Yu.E., Master's student LU LEI

**National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”**

THE DESIGN OF NETWORK SECURITY LOG ANALYSIS SYSTEM BASED ON ELK

Анотація

К.т.н., доцент Боярінова Ю.Є., магістрант Лу Лей

Проектування системи аналізу журналів мережевої безпеки на основі ELK

В статті розглянуто зіставлення сигнатур відомих загроз та аналіз поведінки невідомих загроз. На основі компонентів з відкритим кодом ELK та Kafka створено систему аналізу журналів безпеки. Запропоновано політику безпеки та використання алгоритмів машинного навчання для виявлення загрози та аномальної поведінки програмного забезпечення, що дозволить своєчасно вживати заходів безпеки.

Introduction

Today, as the application of internet technology in businesses becomes increasingly complex, the growing number of cyberattacks poses a significant threat to businesses. Large enterprises face challenges in securely managing complex computer environments and the data transmitted across networks, determining appropriate security policies and levels, and ensuring they can effectively meet business needs. The key to resolving these issues lies in identifying the threats. While defending against all attacks is difficult, rapidly locating and responding to them is essential.

Cybersecurity threats are categorized as external and internal. External threats typically involve hacker activity, including virus distribution via email, phishing, malicious scanning to exploit vulnerabilities, and DDoS attacks, all aimed at stealing and damaging company data. Internal threats typically involve malicious activities such as theft, misuse, or corruption of confidential data, as well as abnormal behavior caused by improper operation, such as misconfiguration and resource consumption. While external threats from attackers are obvious, the damage from within an organization is more subtle and its extent is difficult to assess. This is because internal users are not only more vulnerable to system intrusion but also possess knowledge of the location of critical data. After identifying these threats and risks, organizations must develop appropriate security strategies, establish comprehensive barriers, and establish effective security systems to detect and respond to these threats in real time[1,2].

Problem statement

The log analysis technology studied in this article collects node logs in real time to monitor network traffic, terminal devices, and application status. This helps security researchers quickly identify threats and abnormal behavior before or during security incidents occur, allowing them to take timely response measures.

Terms, abbreviations

ES - Elasticsearch

ELK stands for Elasticsearch, Logstash, and Kibana, three components of the Elasticsearch stack[3].

Log Analysis Technology

The first technology is ELK, which include three components of the Elasticsearch stack. These three are core components, but they are not the only ones. Figure 1 shows a typical ELK system architecture[4].

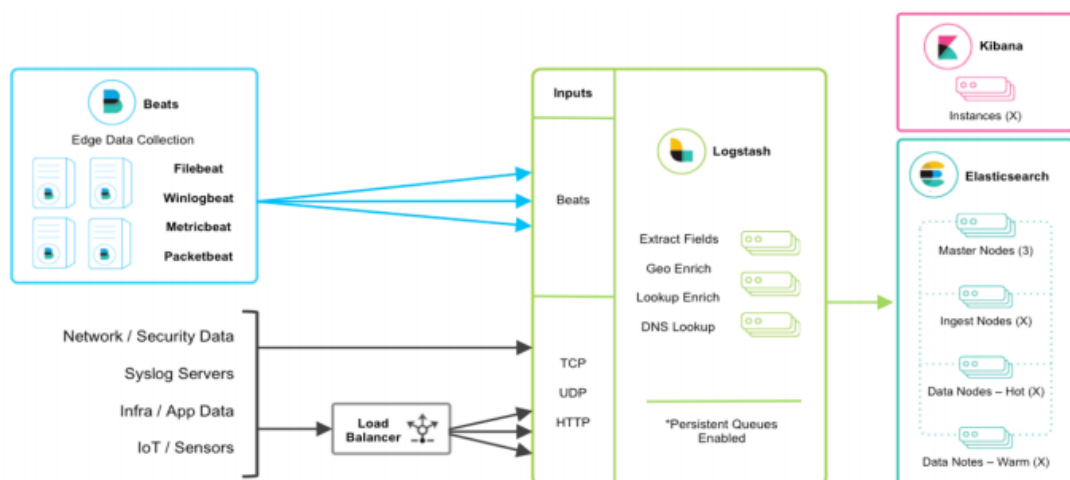


Fig.1. ELK architecture

The second technology is Logstash. It is an open-source data collection engine with real-time pipeline capabilities, dynamically unifying data from disparate sources for a variety of advanced downstream analytics and visualization use cases.

Logstash uses separate threads to implement the input, filtering, and output modules. Data is exchanged between threads in the form of events. In production, the Logstash process is divided into two distinct roles. The Shipper, a log collector, runs on the application server and is solely responsible for reading and forwarding data, minimizing operational load. It also monitors changes in local log files. The Indexer, a log storer, runs on a separate server, parsing data and writing it to Elasticsearch or local files. The Broker is the log hub, connecting multiple

Shippers and Indexers. As a stateless application, Logstash, combined with a message queue, facilitates linear scalability.

The next technology is Kibana. It is open-source visual data analysis software that allows you to search and display data in Elasticsearch, using charts, tables, and maps for diverse data analysis and presentation. Initially, you load data from Elasticsearch into Kibana using index patterns. Kibana's user interface is somewhat similar to Splunk. Kibana has three main functions: search, visualization, and dashboards.

The next technology is Beats. Beats is an open-source agent from Elasticsearch that collects system monitoring data. It's a general term for data collectors that run as clients on the monitored servers. It can send data directly to Elasticsearch or via Logstash for subsequent data analysis activities.

In a production environment, when transferring logs directly from each Logstash/Beat server to Elasticsearch, the following factors can affect performance. First, excessive client connections place an additional burden on Elasticsearch. Second, network jitter can affect the collection process, which in turn affects the production environment. Third, deploying collection components on production servers can hog resources. Therefore, the system introduces a message queue to address these issues. Kafka is a high-throughput, low-latency, and high-concurrency distributed publish-subscribe log service[5].

Design and Implementation of Log Analysis System

The security log analysis system consists of four modules: log acquisition module, data processing and storage module, log indexing module, and visualization and analysis module.

1) Log Collection. Logs are collected using the Beat component; logs from security applications are collected using Filebeat; network traffic is collected using Packetbeat.

2) Message Queue. Kafka acts as a data pipeline, connecting the log collection module and the data processing module.

3) Log Processing. Since there is no unified standard for log formats across applications, logs need to be formatted before indexing, and necessary additional information, such as collection timestamps and source servers, needs to be added. Logs with clear formats are processed using ES ingest nodes, while logs with complex formats are processed using Logstash.

4) Log Storage. Data from the last three days is stored in ES data nodes; data from three days to one year is imported from ES nodes into HDFS storage.

5) Log Indexing. Logs are imported into Elasticsearch, and the index is created all at once during the import process.

6) Visualization Interface. Users can execute queries on the Kibana interface to obtain results, export reports, or use machine learning algorithms to analyze the

results.

Example 1 .Detecting the WannaCry virus

WannaCry is a worm-like ransomware that spreads by exploiting the "EternalBlue" vulnerability.

Identify Malicious Network Activity Based on IPS Signatures

Detect malicious network activity initiated by attackers by matching network traffic behavior, identifying the source/destination IP addresses that trigger relevant signatures.

Discover Potential Malicious Website Access Using WSA Scans

Use WSA network scanning to discover infected hosts connecting to malicious domains, and prevent potential worm spread by changing routes (black hole routing).

Obtain Infected Hosts from AMP Alert Logs

Enterprise computers have AMP software installed. A field name "detection" in the AMP log containing the keyword "WannaCry" indicates that the host has been infected.

Obtain Infected Hosts from DNS Logs

Drowning refers to DNS servers preventing access to specific domains by providing error messages. By querying DNS drowning logs, the infection of internal enterprise hosts with the WannaCry virus can be quickly detected.

Example 2. Research on Log Analysis Based on Machine Learning Algorithms

The K-Nearest Neighbors (KNN) algorithm is a classic classification algorithm, suitable for samples with significant overlap.

The Naive Bayes classifier (NB) classification algorithm is suitable for samples with independent attributes.

The Long Short-Term Memory network (LSTM) algorithm is well-suited for time-series mining deep learning.

Algorithm detection process:

- 1) Select a suitable log data source, export some logs that meet the criteria, and filter the fields of interest.
- 2) Divide the dataset into a training set and a test set. The training set needs to be manually labeled.
- 3) Abstract the features of the detection target, vectorize them, and obtain the feature vector.
- 4) Train the model using the training set.
- 5) Use the test set as the model input, perform cross-validation, and calculate the accuracy.

Tests have shown that this experiment, through training, identifies the

operating habits of internal users within an enterprise and detects internal attacks by attackers posing as internal users.

Conclusions

This paper studies the features of the log analysis suite ELK, and discusses the optimal system architecture and performance solution by combining it with the message queue Kafka and distributed deployment. An enterprise-level log analysis system is designed, consisting of four modules: log collection, log processing and storage, log indexing, and visualization. Mechanisms such as distributed deployment, redundancy backup, and hot/cold data separation are included to ensure data consistency and reliability.

For known attacks, this paper selects three common scenarios, defines several security policies, and executes detection. Results show that the security policies can quickly and effectively detect viruses and attacks with known behavioral characteristics and identify infected hosts.

For unknown attacks, this paper selects three algorithms, including classification, clustering, and deep learning techniques. Using formatted logs from the secure log analysis system as input, the algorithms are learned and tested in different scenarios by extracting and statistically analyzing features. Cross-validation shows that the detection accuracy of all algorithms reaches 90%.

References

1. Nabil M, Soukainat S, Lakbabi A, et al. SIEM selection criteria for an efficient contextual security[C]//2017 International Symposium on Networks, Computers and Communications (ISNCC). IEEE, 2017: 1-6.
2. Kumar A, Bandyopadhyay A, Bhoomika H, et al. Analysis of network traffic and security through log aggregation[J]. International Journal of Computer Science and Information Security (IJCSIS), 2018, 16(6).
3. Live logs and prosper, on-line :<https://www.elastic.co>
4. Son S J, Kwon Y. Performance of ELK stack and commercial system in security log analysis[C]//2017 IEEE 13th Malaysia International Conference on Communications (MICC). IEEE, 2017: 187-190.
5. Zhou Minfei. Design and implementation of real-time log stream processing system based on Kafka and Storm[D]. Guiyang: Guizhou University, 2017.