

Data collection, aggregation, and analysis in the Internet of Everything environment – a case study

Andrzej Gołabek
0000-0001-7973-9534
InfoSoftware Polska Sp. z o. o.
ul. Przemysłowa 5A
38-457 Szczepańcowa, Poland
Email:
agolabek@infosoftware.pl

Piotr Sowa,
Mateusz Kozik
InfoSoftware Polska Sp. z o. o.
ul. Przemysłowa 5A
38-457 Szczepańcowa, Poland
Email:
psowa@infosoftware.pl,
mkozik@infosoftware.pl

Andrzej Paszkiewicz,
Marek Bolanowski
0000-0001-7573-3856
0000-0003-4645-967X
Rzeszow University of
Technology,
Al. Powstańców Warszawy 12,
35-029 Rzeszów, Poland
Email:
apaszkiewicz@infosoftware.pl,
mbolanowski@infosoftware.pl

Abstract— This article presents a practical approach to topics related to the Internet of Everything (IoE) environment, specifically describing the main mechanisms and challenges associated with the collection, aggregation, and analysis of data from multiple sources. In distributed IoE environments, data is transmitted using various communication protocols and at different frequencies. Both network packets and logs transmitted by IoE devices are specific to a given device type and often vary depending on the manufacturer. In large-scale IoE environments, the very large volume of data transmitted in various formats requires the implementation of effective mechanisms to aggregate it in near real time, especially when considering its analysis in the context of cybersecurity. This article presents a general operational diagram of a system enabling the analysis of data from a distributed IoE environment, incorporating mechanisms for data acquisition from multiple sources, caching, scaling, and multi-stage data analysis that machine learning and artificial intelligence support.

I. INTRODUCTION

INTERNET of Everything (IoE) environments represent an extension of the traditional concept of the Internet of Things (IoT), encompassing not only end devices but also users, business processes, applications, analytical systems and network infrastructure [1, 2, 3]. In an IoE environment, data is not generated solely by individual sensors or telemetry devices, but by an entire ecosystem of interconnected elements. In practice, IoE environments generate vast amounts of data from many heterogeneous sources [4, 5]. This data includes, amongst other things, system logs, telemetry data, network packets, diagnostic messages, security events, alerts from IDS/IPS systems, syslog data, NetFlow, and information from industrial devices, OT networks and cloud infrastructure. This data is transmitted using various mechanisms and

communication protocols, such as MQTT, Modbus RTU, Syslog, HTTP/HTTPS, UDP, RTP, REST API, CoAP and NetFlow, amongst others. Furthermore, many manufacturers of the same types of devices use different formats for transmitted data and messages, apply varying levels of log detail, and employ different data transmission frequencies, which necessitates the use of additional mechanisms to normalize this data prior to further analysis [6, 7].

From the perspective of customers and organizations implementing IoE solutions, it is essential to ensure an effective, consistent and, above all, secure method of collecting and aggregating data from numerous dispersed locations. Once this data has been appropriately filtered, normalized, classified and, in many cases, enriched with additional context, it should be accessible via a central analytics system. The collected and aggregated data form the basis for further analysis, particularly about the correct operation and performance of the entire IoE environment, as well as cybersecurity. Therefore, in IoE environments, it is essential to implement scalable systems that enable the collection, aggregation and analysis of data in near real time. Furthermore, the systems implemented must be equipped with mechanisms to mitigate the risk of data loss and must ensure business continuity even under increased load [8, 9].

Despite the growing number of solutions designed to monitor and secure IoE environments, there remains significant scope for research and implementation regarding their practical application in small and medium-sized enterprises and local government bodies. The SIEM, SOAR, IDS/IPS systems and stream analysis platforms available on the market are designed for large organizations that possess technical and financial resources, as well as specialized staff. For small organizations, implementing such solutions is often impossible

due to their high costs and the lack of access to specialized staff with experience in solutions integrating multiple heterogeneous data sources [10, 11]. At the same time, these organizations use various IT systems (often distributed), IoT/IoE devices, cloud applications and security tools from different suppliers in their day-to-day operations, which creates a need for flexible analytical solutions tailored to the organization's actual capabilities. A particular challenge remains the acquisition, normalization and correlation of data generated across the entire IoE environment implemented within the organization. The diversity of communication protocols, data formats and transmission frequencies means that small and medium-sized enterprises require simplified, modular and scalable systems supporting near-real-time data analysis [12].

The innovative aspect of the approach presented in this article lies in the development of a system concept for the IoE environment, combining mechanisms for data aggregation, queuing, normalization, stream analysis, anomaly detection and artificial intelligence support in cybersecurity processes [13, 14]. A key element of the proposed approach's innovation is the use of a Feedback Loop mechanism, which creates feedback between the analytical layer and the data pre-processing layer. This mechanism allows for the automatic updating of correlation rules, the modification of stream analysis parameters, and the adjustment of feature vectors used in anomaly detection and cybersecurity analysis. As a result, the system can respond not only to detected incidents but also to changes in traffic characteristics, the location of data sources, load levels, and threat profiles. This approach transforms the IoE environment from a passive data aggregation and monitoring system into an adaptive architecture capable of self-adjusting to changing operational conditions.

From the perspective of small and medium-sized enterprises, modular architecture allows the system's functional scope to be tailored to the organization's needs and capabilities, without the need to build costly infrastructure. This approach enables the gradual implementation of monitoring, event correlation, anomaly detection and automated incident response functions, which significantly lowers the implementation barrier and enhances their level of cybersecurity [15].

II. CUSTOMER REQUIREMENTS, LOG CHARACTERISTICS, AND SECURITY MECHANISMS IN IOE ENVIRONMENTS

The main requirements that customers place on IoE environments are, above all, high system availability and scalability, reliable data transmission, the ability to centrally monitor infrastructure, and ease of integration with existing IT solutions. Scalability is of particular importance, as the number of devices, sensors, applications and data-generating systems is growing rapidly as organizations expand. Another key requirement is the ability to analyze data in near real time, particularly in environments with limited hardware resources, which is what small and medium-sized enterprises can afford. It is important to customers that such solutions enable the rapid detection of failures, anomalies, configuration errors and potential security incidents. Automation is crucial, and

more specifically the implementation of artificial intelligence and machine learning mechanisms that will support customers through the automatic classification of events, the detection of unusual communication patterns, and the reduction of false alarms.

The information logged by IoE devices is highly varied. It includes telemetry data, system events, diagnostic information, network traffic data, error messages, security alerts and information on the status of end devices. The log formats, transmission frequency and level of detail depend not only on the type of device, but also on the manufacturers. Some IoE devices transmit data periodically, whilst others, such as sensors, only send logs after a change in status or following a specific event. Therefore, one of the key stages in processing data from such a diverse environment is its normalization and standardization. Standardizing the data enables its subsequent analysis and correlation. This process includes, amongst other things, parsing messages, assigning timestamps, classifying event types, and adding context such as the device's location, type, IP address or severity level.

In IoE environments under development, data security is of paramount importance alongside effective data collection and aggregation mechanisms, and customers are increasingly drawing attention to this. The basic security measures that must be implemented in such environments include data transmission encryption, network segmentation, access control, multi-factor authentication (2FA/MFA), device authentication, and monitoring of the integrity of transmitted information. The principles of Zero Trust architecture are also frequently applied, in which no element of the IoE infrastructure can be considered trusted by default. Every device, user and application should be verified before being granted access to system resources. SIEM and SOAR systems also play a significant role, enabling centralized log collection, event correlation, and the identification of and automatic response to potential threats. The combination of effective and secure mechanisms for log collection and aggregation, correlation rules, AI-powered analysis, and SOAR playbooks create an IoE environment with a high level of cybersecurity.

III. CASE STUDY AND KEY COMPONENTS OF THE IOE ENVIRONMENT

A sample Internet of Everything (IoE) environment designed to collect, aggregate and analyze data from multiple distributed data sources is presented as a case study. A diagram of the proposed environment is shown in Fig. 1. This diagram includes customer requirements relating to scalability, high availability, fault tolerance, near-real-time data analysis and security-related elements.

The first element shown in the diagram is the data sources, which include IoE devices, telemetry systems, network infrastructure and security components such as EDR, IDS/IPS and UTM. The data is transmitted using various communication protocols, such as MQTT, Syslog, HTTPS, NetFlow and Modbus, to so-called data collectors and data producers.

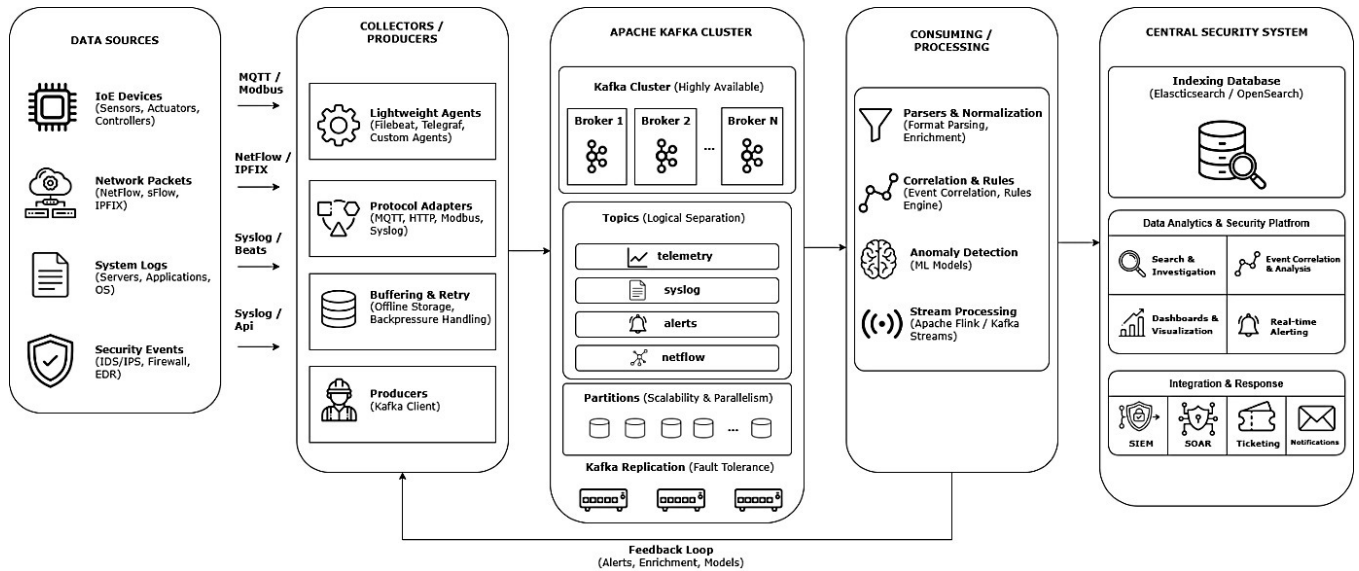


Fig. 1. Diagram of an example environment for collecting, aggregating, and analyzing data from multiple IoE sources.

The collector and producer layer is primarily responsible for collecting data, performing initial filtering, and normalizing it. This component of the environment includes, amongst other things, log-collecting agents, communication protocol adapters that enable the integration of devices from different manufacturers, and data producers such as the Kafka Client.

Another component of this environment is the Apache Kafka cluster, which acts as a distributed queuing and data transport system. The data transmitted is logically divided into topics relating to telemetry, syslog data, alerts and Net-Flow. This layer utilizes partitioning and replication mechanisms that ensure high performance, fault tolerance and the ability to process very large volumes of data in parallel in near real time.

The fourth element shown in the diagram is the data processing and analysis layer. It is primarily responsible for parsing incoming logs, event correlation, stream analysis and anomaly detection. Machine learning algorithms have been implemented within this layer to support the initial identification of security incidents and unusual communication patterns.

The final layer of this environment is a central security system that utilizes an indexed database, enabling rapid searching, analysis and correlation of events. This system integrates SIEM and SOAR solutions, as well as a ticketing system, to support the SOC team in responding to security incidents. A key feature of this system is the ability to map events to the tactics and techniques described in the MITRE ATT&CK matrix. This allows detected incidents to be classified not only as individual alerts, but as elements of a potential attack chain. The system also utilizes Indicators of Compromise (IoC), such as suspicious IP addresses, domains, file extensions, unusual process IDs, malware signatures or characteristic patterns of network communication, to automatically enrich logs and compare current events against databases of known

threats. In addition, the central security system collects data on software versions and device types and compares this with

CVE (Common Vulnerabilities and Exposures) databases to identify potential attack vectors. This layer also incorporates artificial intelligence mechanisms in the form of machine learning models that analyze data from a behavioral perspective (UEBA) and an AI assistant that supports the SOC team in analyzing incidents, issues and detected anomalies. The implemented artificial intelligence significantly streamlines the response process to detected anomalies and security incidents and provides suggestions regarding correlation rule settings and SOAR scenarios.

A key component of this environment is the Feedback Loop mechanism, which is responsible for transmitting feedback between the analytical layer and the data pre-processing layer (collectors). This mechanism provides feedback that enables the system to dynamically adapt its operation based on the results of ongoing analysis of logs and security events. It allows for the automatic updating of correlation rules both in time and space, the modification of stream analysis parameters, and the adjustment of feature vectors for more accurate cybersecurity analysis. Adaptation over time means that the system can respond to the changing dynamics of the IoE environment, such as an increase in the number of messages, the emergence of anomalies, periodic overloads, changes in communication patterns, or the occurrence of a security incident. From this perspective, correlation rules, detection thresholds and stream analysis parameters are not static, but can be adjusted based on the current state of the system and observed events. Spatial adaptation, on the other hand, refers to the ability to vary the method of data analysis and processing depending on the source's location, network segment, device type, the role of a given component in the architecture, and the criticality level of the monitored area. Different rules may apply to control systems, different ones to stream sources, and yet

others to application data sources or telemetry sensors. As a result, the IoE environment is not treated as a uniform communication space, but as a collection of areas with varying levels of risk, differing operational significance, and distinct requirements regarding security and performance.

IV. SUMMARY

The rapid growth of Internet of Everything environments is leading to a steady increase in the number of devices, applications and services that generate data continuously and in a distributed manner. In practice, this means there is a need to build modern, scalable systems capable of efficiently collecting, aggregating and analyzing large volumes of data from heterogeneous sources in near real time, whilst operating with severely limited hardware resources.

The Feedback Loop mechanism used is one of the key elements of the innovative architecture proposed in this article. This mechanism enables the transfer of analysis results from the analytical layer to the layer of collectors and data pre-processing components, allowing for the dynamic modification of correlation rules, stream analysis parameters, and feature sets used in detection processes. In practice, this means the system can be continuously adapted to the current threat profile, infrastructure load, and the significance of individual data sources.

The proposed solution is not limited to the passive collection and analysis of information but creates an adaptive mechanism that supports the automation of cybersecurity, the reduction of false alarms, and a more efficient use of IoE infrastructure resources. The introduction of feedback between the analysis and data acquisition layers increases the resilience and scalability of the entire environment. It also allows for the further development of the system towards self-learning architecture, in which decisions regarding prioritization, filtering and correlation of data are gradually optimized based on observed events, anomalies and security incidents.

Of particular importance is the ability to adapt the system's operation in both temporal and spatial dimensions. The temporal dimension refers to the system's ability to respond to changes occurring over time, such as an increase in traffic volume, the occurrence of anomalies, changes in communication patterns, infrastructure overload, or the emergence of a new attack pattern. This allows detection and correlation rules to be updated not only periodically, but also in response to current events observed in the environment. The spatial dimension, on the other hand, concerns the ability to differentiate analysis depending on the location of data within the IoE architecture. The system can apply different detection

thresholds, correlation rules and levels of analytical detail for different network segments, physical locations, device classes and areas with varying levels of criticality. This avoids treating the entire infrastructure uniformly and allows analytical resources to be concentrated where the risk or operational significance of the data is greatest.

REFERENCES

- [1] E. Siow, T. Tiropanis, W. Hall, "Analytics for the Internet of Things: A Survey", *ACM Computing Surveys*, vol. 51, no. 4, pp. 1–36, 2018. DOI: 10.1145/3204947.
- [2] M. Ge, H. Bangui, B. Buhnova, "Big Data for Internet of Things: A Survey", *Future Generation Computer Systems*, vol. 87, pp. 601–614, 2018. DOI: 10.1016/j.future.2018.04.053.
- [3] Y. Hajjaji, W. Boulila, I. R. Farah, I. Romdhani, A. Hussain, "Big Data and IoT-Based Applications in Smart Environments: A Systematic Review", *Computer Science Review*, vol. 39, 100318, 2021. DOI: 10.1016/j.cosrev.2020.100318.
- [4] D. Fawzy, S. M. Moussa, N. L. Badr, "The Internet of Things and Architectures of Big Data Analytics: Challenges of Intersection at Different Domains", *IEEE Access*, vol. 10, pp. 4969–4992, 2022. DOI: 10.1109/ACCESS.2022.3140409.
- [5] A. Naghib, N. Jafari Navimipour, M. Hosseinzadeh, A. Sharifi, "A Comprehensive and Systematic Literature Review on the Big Data Management Techniques in the Internet of Things", *Wireless Networks*, vol. 29, pp. 1085–1144, 2023. DOI: 10.1007/s11276-022-03177-5.
- [6] I. Ali, S. Sabir, E. Khan, "Privacy-Preserving Data Aggregation in Resource-Constrained Sensor Nodes in Internet of Things: A Review", *Future Computing and Informatics Journal*, vol. 3, no. 1, pp. 41–50, 2018. DOI: 10.1016/j.fcij.2017.11.
- [7] J. Hou, L. Qu, W. Shi, "A Survey on Internet of Things Security from Data Perspectives", *Computer Networks*, vol. 148, pp. 295–306, 2019. DOI: 10.1016/j.comnet.2018.11.026.
- [8] G. Calderon, G. del Campo, E. Saavedra, A. Santamaria, "Monitoring Framework for the Performance Evaluation of an IoT Platform with Elasticsearch and Apache Kafka", *Information Systems Frontiers*, vol. 26, pp. 2373–2389, 2024. DOI: 10.1007/s10796-023-10409-2.
- [9] T. P. Raptis, A. Passarella, "A Survey on Networked Data Streaming with Apache Kafka", *IEEE Access*, vol. 11, pp. 85333–85350, 2023. DOI: 10.1109/ACCESS.2023.3303810.
- [10] A. Emer, M. Unterhofer, E. Rauch, "A Cybersecurity Assessment Model for Small and Medium-Sized Enterprises", *IEEE Engineering Management Review*, vol. 49, no. 2, pp. 98–109, 2021. DOI: 10.1109/EMR.2021.3078077.
- [11] N. Rawindaran, A. Jayal, E. Prakash, "Machine Learning Cybersecurity Adoption in Small and Medium Enterprises in Developed Countries", *Computers*, vol. 10, no. 11, 150, 2021. DOI:10.3390/computers10110150.
- [12] M. Ulrich, F. Frank, A. S. Klug, "An AI Adoption Model for SMEs: A Conceptual Framework", *IFAC-PapersOnLine*, vol. 54, no. 1, pp. 702–708, 2021. DOI: 10.1016/j.ifacol.2021.08.082.
- [13] A. Chatterjee, B. S. Ahmed, "IoT Anomaly Detection Methods and Applications: A Survey", *Internet of Things*, vol. 19, 100568, 2022. DOI: 10.1016/j.iot.2022.100568.
- [14] K. DeMedeiros, A. Hendawi, M. Alvarez, "A Survey of AI-Based Anomaly Detection in IoT and Sensor Networks", *Sensors*, vol. 23, no. 3, 1352, 2023. DOI: 10.3390/s23031352.
- [15] J. Kinyua, L. Awuah, "AI/ML in Security Orchestration, Automation and Response: Future Research Directions", *Intelligent Automation & Soft Computing*, vol. 28, no. 2, pp. 527–545, 2021. DOI: 10.32604/iasc.2021.016240.