ScienceSoft
Professional Software Development

Supporting New Data Sources in SIEM Solutions: Key Challenges and How to Deal with Them

White Paper

Serguei Tchesnokov — Tchesnokov@scnsoft.com
Alexei Zhurba — AlexeiZhurba@scnsoft.com
Uladzimir Radkevitch — Radkevitch@scnsoft.com
Aliaksandr Jurabayeu — Ajurabayeu@scnsoft.com

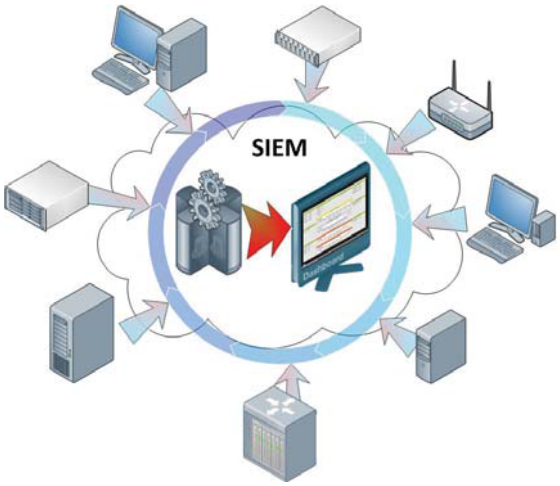
ABSTRACT

Security Information and Event Manager (SIEM) is a merger of two previously separated product types: Security Information Manager (SIM) and Security Event Manager (SEM). SIEM provides real–time comprehensive analysis of security audit–related data gathered from various data sources, such as operating systems, applications, network devices etc. Analyzed data is presented in the form of reports for compliance and threat management purposes.

The use of SIEM solutions is undergoing rapid growth. According to Gartner research, the number of inquiry calls from SIEM end users has been growing by 20–35% and even more each year. In 2010, the security software market grew more than 11%, exceeding $16.5 billion, 20% of which was a share of SIEM products alone. However, their implementation can be costly and difficult due to the complexity involved in handling data sources. This white paper identifies key challenges in the development of solutions for data sources and provides insights into how to deal with these challenges. In particular, we address the selection of sources of audit log data, selection of supported audit events, identification of sources for missing data, original logs concept and globalization issues.

INTRODUCTION

SIEM users face the pressure of high costs related to the selection of a SIEM product, its deployment, integration and day–to–day use. Such cost pressure stems from substantial initial cost of the product, additional expenses for specific hardware and third–party product licenses, cost of external consultants, staff training, operational and maintenance expenses etc. In the end, what was initially considered as an affordable price for the chosen SIEM solution, may eventually turn out to be a large hole in the budget with an enormous total cost of ownership.



Lack of connectivity with the log data sources is one of the key impediments to successful implementation of SIEM as well as a significant cost driver. Such connectivity is usually provided via adapters (also known as “connectors”, “collectors”, “Event Sources”, etc.), each of which provides support for a single source of log data even if the source is very generic like Syslog or SNMP. Since there are thousands of potential sources of data that use endless varieties of log formats and structure, no out–of–the–box SIEM product can provide support to all such sources. This is due to the high cost of development, support and technical difficulties associated with the creation of adapters. Software vendors do not tend to migrate their logging to universal frameworks, even if such frameworks do exist and may fit their needs. Almost all

software vendors use their own logging solutions, which significantly differ from manufacturer to manufacturer.

SIEM vendors aim to provide frameworks and universal connectors that help the user and vendor's own staff to develop custom log data sources solutions. However, more often than not, plugging in the missing log data sources becomes the problem for the end user and needs to be solved by their own personnel, or external consultants acting on their behalf.

The functionality provided by a SIEM solution is not the only part ensuring successful implementation of a log data source connector. The other important factor is a strong knowledge of security audit and logging mechanisms, and deep understanding of processed log data.

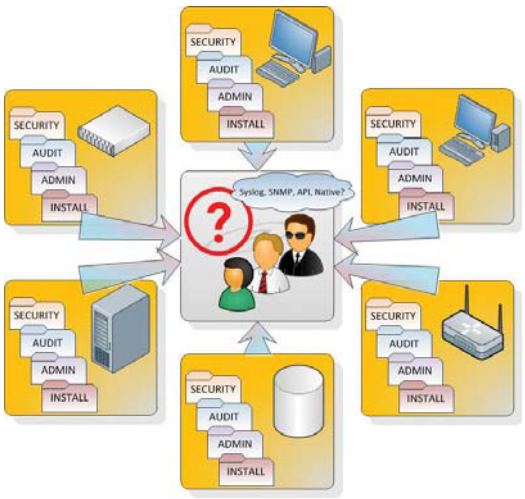
MAJOR CHALLENGES IN SUPPORTING THE LOG DATA SOURCE

Regardless of the SIEM solution considered, the main implementation challenges are almost all the same. These challenges result from the complexity of software auditing and logging subsystems, lack of technical documentation for such software and lack of experienced personnel. Technical knowledge about the software and experience in SIEM solutions development are far less important issues compared to the implementation specifics of many software manufacturers. Since the software vendors do not follow any universal standards in logging solutions, the following problems are likely to appear and must be taken into account:

1. Selection of sources of audit log data.

Quite often a particular software or device provides more than one type or source of audit log data. There can be more than one audit log, the ability to switch between native logging and Syslog or SNMP, different logging facilities may exist within one and the same software product or device, or logs can be stored in different places and serve different needs. Since the SIEM implementation has a clearly defined aim, all available sources of audit log data must be examined thoroughly and the necessary ones must be considered for processing.

Today it is no surprise if software generates a few gigabytes of audit log data per day. Indeed, some applications can even generate a few gigabytes of progress log in less than 1 hour. Since the total amount of logs can be quite large, it is important to select an appropriate source of audit log data. It is also necessary to keep in mind that different types of log can provide different amounts of relevant and useful information. The most detailed and comprehensive log is not necessarily the best choice for SIEM processing. Defining the best log is subject to information security standards, policies and regulations applied.



Since SIEM vendors cannot provide support for all existing log data sources, they usually provide generic support for common logging mechanisms such as Syslog, SNMP, Windows Event Log, etc. So, the more standardized the source selected, the more likely it is supported by a particular SIEM software product. In some

cases, the selection of appropriate log data source saves time and helps avoid such problems as missing and duplicated events.

2. Selection of audit events to support.

Software vendors do not create log data sources that cater specifically for the requirements of SIEM solutions. Depending on the requirements of a specific SIEM solution it may be necessary to obtain and process information from several logs, filter out the unnecessary event types from a log data source, or take other actions to limit the amount of data collected with SIEM software.

In general, no information security standard or regulation requires collection of all available log data. Typically, they only require the collection and retention of audit log data related to certain information assets and activities. Thus, the event filtering is not just allowed by security standards, but strongly recommended from the operational point of view.

To filter particular event types as non–relevant, it is necessary to understand what they actually mean. However, filtering is not the only reason to investigate and understand events. Another good reason to know the meaning and detailed structure of events is their direct use as a source of information for SIEM reporting.

Since the SIEM software is expected to provide a lot of smart reporting, the log data is not only aggregated and parsed, but also processed in many different ways. Almost all modern SIEM products normalize event data, making them more suitable for further analysis. Processing and normalization techniques vary from product to product, presenting the log data in different ways and according to different models but the expected outcome is always the same:

— Similar events from different operating systems, devices, or products must be clearly and unambiguously described by one single normalization term;

— Normalized event representation must answer the question “Who did What, When, and Where?”

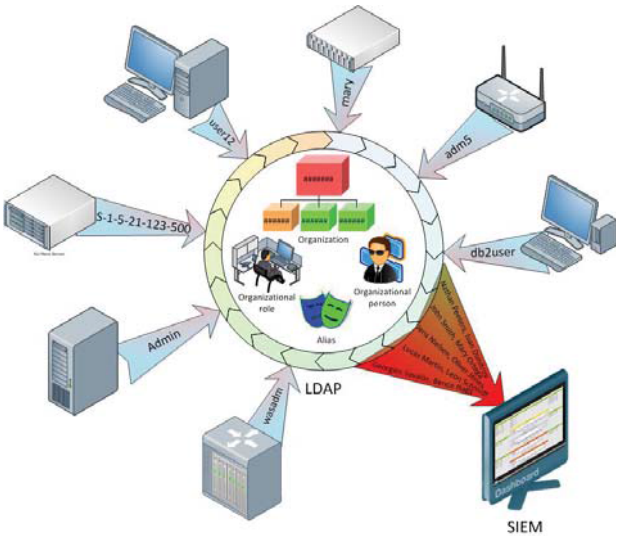
3. Missing data.

Since software vendors do not care about SIEM implications of their log files, there is a high probability of coming across log data that do not contain all the details required for a proper SIEM analysis. Missing details do not necessarily make an event useless for the SIEM analysis, but it really depends what details are missing. While missing an actor name (Who) may be acceptable, missing a timestamp (When) makes an event almost useless for the proper SIEM analysis.

In some cases missing data can be obtained from additional sources of information such as user directories, databases, configuration files, etc. Being available in some form, such

additional data just needs to be applied (manually or automatically) to collected event log data. Depending on the nature of missing data, it may be necessary to collect additional information each time the event data is being collected; at regular intervals or just once during a significant period of time.

Clarification of missing details is not the only situation when additional data can be applied. Even if an event record contains all the required information, some forms of additional data may be valuable during the event analysis. An example of such additional information is a correspondence between user names and SIDs, in case event records contain SIDs and not the user names. Another good example is information about groups that a particular user belongs to.

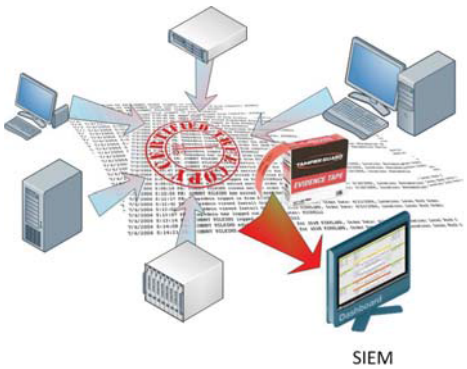


Processing sources of additional information out of the box is an extremely common practice for all modern SIEM products. In most difficult cases, however, such additional information can be applied manually through the creation of specific reports and rules.

4. Retention of original logs.

One of the important log management functions available in SIEM software is retention of original logs. Information security standards and regulations require log retention without providing explanation of what the original log is and how to ensure the log's originality.

Depending on the particular case, it may be necessary to distinguish original (raw) log data sources and those extracted from raw and converted to human–readable text dumps (CSV, XML, etc.). The data extracted from raw log to human–readable format does not necessarily contain exactly the same set of information as its origin – it is likely that the data is enhanced with additional details, human–readable representation of internal terms, etc.



However, most important is not how a particular log is generated, but the clear and unambiguous tracking of all the changes made to its event records. The activities requiring special attention are modification or deletion of existing records, creation of new events or event details, modification of log file properties, etc. Tracking logs integrity is what is really needed to meet requirements imposed by security standards, like ISO 27001, PCI DSS, HIPAA, GLBA, BASEL, COBIT, SOX and others.

5. Globalization.

Obtaining audit trails from various log data sources, and processing and analysis of gathered data are the basic features of any SIEM product. Although all these tasks do not seem significantly complex at first, the situation changes dramatically in relation to the National Language Support (NLS). Relatively small software manufacturers, which target local businesses, do not provide globalization for their products at all. By contrast, large and well–established software manufacturers that cater for international businesses across the globe, use special logging frameworks to provide information using several national languages. Depending on the number of supported languages, the cost of translations can have a significant impact on the overall price of SIEM implementation. Moreover, collecting, processing and analysis of NLS–enabled audit logs also pushes the envelope for SIEM developers in terms of using various techniques and technologies. The following is a more detailed list of most common NLS issues that have to be considered by SIEM developers:

The Syntax Problem

Location of relevant pieces of collected information depends significantly on the language environment they come from. One sentence can be constructed differently in different languages and this leads to a 'location' problem. The worst case scenario is when the SIEM development team does not have the language expertise and it involves the assistance of native language experts and requires access to target product information.

The Semantics Problem

Data written in the log depends significantly on the linguistic semantics employed by software manufacturers. That is, different equivalents can be used by different software manufacturers for a single term. This leads to a 'lost in translation' problem. The worst case scenario, when the SIEM development team does not have the correct language expertise, involves the employment of native language experts and requires access to target product information.

The Interface Problem

Executable commands for obtaining the required information may be specific to the language environment being used. The worst case scenario, when the SIEM development team does not have the correct language expertise, involves the employment of native language experts and requires access to target product information.

The Scattered Knowledge Problem

The aforementioned problems require knowledge of particular language syntax and semantics, language–dependent operating system interface and the development process for producing a translated SIEM solution. The following information describes the dependencies, which show how the knowledge is scattered across the different development parties:

a) The manufacturer controls language syntax and semantics, but they are not aware of the challenges that SIEM developers face.

b) SIEM developers control collecting, processing and analysis of data, but depend on a particular language syntax and semantics used by the software manufacturer.

c) Language experts know the language, but they do not know processing and analysis of data.

In almost all of the cases, none of the above can solely provide a suitable translated SIEM

solution. The worst case scenario, when the SIEM development team does not have the correct language expertise, involves the employment of native language experts and access to target product information, and requires various discussions with the manufacturer.

The High Cost Problem

Compared to English language–only implementations, NLS support introduces up to 50% of the development effort. Moreover, considering the support for all possible languages that a particular software implementation may provide, the translation per se may take a long time and will have a tremendous impact on the cost and duration of the project. This ratio between the cost of globalization and the cost of original development is far greater than for typical applications that do not need to process language–dependent input data.

CONCLUSION

The implementation of additional log data sources is not easy or straightforward for any SIEM solutions. It requires a wide range of technical skills and expertise, such as general system and software administrator experience, deep knowledge of auditing and logging subsystems, SIEM product area expertise and software development skills, including knowledge of programming languages. In some cases even in–depth expertise does not protect the implementer from problems such as lack of information, or incorrect or misleading information about 3rd party product logs. Overcoming all these challenges requires a certain amount of time, which, for the majority of projects, is the most important factor. Training of personnel costs time, whereas external experts are expensive to hire even for small and sporadic projects. In the end, everything relies on careful planning and a fine balance between the cost of development and project duration.

Luckily, the development of a particular SIEM solution can be controlled with a systematic approach to these challenges. Determination of sources of audit data as well as appropriate selection of audit events to support and globalization support analysis in early stages can substantially reduce the overall duration and cost of the project. On top of that, identification of sources for missing data, following original log concept and globalization imperatives will provide a better design and compliance with security standards, policies and regulations. A better design and suitable implementation, in turn, will ensure better maintainability as well as efficient support in the future.

CONTACT INFORMATION

SCIENCESOFT USA

5900 S. Lake Forest Dr., Suite 300
McKinney, TX 75070, USA
Phone: +1 214 306 68 37
Email: contact@scnsoft.com
Web: www.scnsoft.com

SCIENCESOFT Finland

Myyrmäenraitti 2
01600 Vantaa, Finland
Phone: +358 92 316 3062
Email: contact@scnsoft.fi
Web: www.scnsoft.com