Data Repository for Security Information and Event Management in Service Infrastructures

Igor Kotenko, Olga Polubelova and Igor Saenko
Laboratory of Computer Security Problems, St. Petersburg Institute for Informatics and Automation (SPIIRAS), 39, 14th Liniya, Saint-Petersburg, Russia

Keywords: Security Repository, Security Information and Event Management, Security Ontology, Data Model, Data Representation, Logical Inference, Service Infrastructure.

Abstract: Design and implementation of the repository is a critical problem in advanced security information and event management (SIEM) systems, which are SIEM systems of service infrastructures. The paper discusses several innovations which are realized to address this challenge. These include the application of an ontological approach for repository data modeling and a hybrid approach to its development, meaning the combined use of relational databases, XML databases and storage of triplets.

1 INTRODUCTION

At present, one of the most important research directions in the area of computer network security is a technology of *security information and events management* (SIEM). The essence of this technology is to ensure coherent boot in a centralized repository of security log records from a variety of sources – "security events", their long- and/or short-term storage, modeling and analysis to detect attacks, generating efficient countermeasures. SIEM technology can make effective safety decisions based on event correlation, data mining, logical inference and data visualization. Using SIEM systems is extremely important to ensure information security of large distributed computer networks, management and financial services of companies as well as for critical infrastructures, such as dams, power plants, etc. (Miller et al., 2011).

Advances of SIEM systems in computer network infrastructures give rise to use such systems in the broader class infrastructures that can be defined as *service infrastructures*. In these infrastructures, in addition to computer networks there are the infrastructures of various types of services (financial, physical, etc.). SIEM systems which can be used in service infrastructures are considered in the paper as *new generation SIEM systems*.

MASSIF (MAnagement of Security information and events in Service InFrastructures) is one of the EU projects, which aims to develop solutions to build a new generation of SIEM systems (MASSIF, 2011). SIEM systems elaborated in MASSIF must have the following new features: removing most of restrictions on the functions imposed by infrastructure; coherent interpretation of the incidents and events at various levels; high degree of reliability and durability in capturing event data; high scalability.

One of the most important components of the SIEM systems used in service infrastructures is a *security repository*, which is a data warehouse that enables to store security information and event data in an internal format and extracts it at the request of other components for identifying security threats and attacks and generating countermeasures.

In SIEM systems of service infrastructures, security event data arrive from a variety of different sources and can be presented in various input formats. A SIEM system produces the normalization of those data and they are converted into an internal format. Then the security data are exposed to correlation analysis (Stevens, 2005). In the SIEM system of new generation it is possible to use the advanced modeling and simulation modules, which also use data stored in the repository to build on their basis the attack and countermeasure graphs (Ingols et al., 2009); (Kotenko et al., 2006).

For these reasons, the main objectives of the repository development are as follows: to design a unified repository, languages and tools for effective management of security information, events and

Kotenko I., Polubelova O. and Saenko I..
Data Repository for Security Information and Event Management in Service Infrastructures .
DOI: 10.5220/0004075303080313
In *Proceedings of the International Conference on Security and Cryptography* (SECRYPT-2012), pages 308-313
ISBN: 978-989-8565-24-2

policies, and logical inference about security; to implement software applications for storing, manipulating, visualizing and validating security information, events and policies based on the unified repository. The paper examines *the main issues of data model design and repository development for new generation SIEM systems*. We could note the following *innovations* that were used to solve this problem. First, for data modeling the *ontological approach* is proposed and implemented. It provides the necessary flexibility of internal data representation in the repository and the possibility of more accurate and high-quality results of queering. Secondly, the *hybrid approach* to implement the repository is suggested. It integrates relational databases, XML databases and stores of triplets. Finally, we propose the advanced *repository architecture* implemented and tested with the data used for attack modeling in SIEM systems.

The rest of the paper is organized as follows. *Section 2* reviews related work in the field of SIEM data processing, representation and storage. Here we consider standards in security event representation, advanced SIEM systems, languages for data representation, and approaches to implement the repository. *Section 3* considers the ontological vulnerability model used in the repository for attack modeling. *Section 4* discusses the issues of the repository implementation and testing based on the data of the SIEM attack modeling module. *Section 5* concludes our results and outlines further research.

2 RELATED WORK

During the analysis of state-of-the-art, we considered perspective and widely used approaches and standards for data representation in area of security information and events management.

Information and event management standards provide the most common rules for representation of security events and incidents. Currently, there are many different standards of security data representation (IDMEF, IODEF, CEE, SCAP, CBE, CEF, XDAS, CIM, etc.). The most popular of them are Common Event Expression (CEE), SCAP (SCAP, 2011), Common Base Event (CBE) (Ogle et al., 2004) and Common Information Model (CIM) (CIM, 2011). For example, *CEE* realizes a comprehensive approach to handling the input stream of information to log management systems, including recommendations to vendors of hardware and software systems that generate the input stream. *SCAP* enables to compile a list of system platforms and applications, set their secure configurations, identify the most critical vulnerabilities, etc.

The main repository solutions in advanced SIEM systems (AlienVault OSSIM, AccelOps, QRadar, Prelude, ArcSight, IBM Tivoli, and Novel Sentinel) are based on relational databases. The storage in OSSIM includes a user-defined, searchable knowledge base of incident solutions (AlienVault, 2011). *AccelOps* SIEM is designed to collect logs generated by Cisco network and security devices, and all the major network vendors' devices. The repository is implemented as online PostgreSQL storage applied for log analysis in real time and for historical log analysis (AccelOps, 2011). *Qradar* stores the entire input event stream to enable detailed forensics and compliance reporting (Miller et al., 2011). *Prelude* (Prelude, 2011) supports three databases: MySQL, PostgreSQL, and SQLlite. *ArcSight Logger 4* collects data in structured and unstructured formats (Shenk, 2009). The system implements role-based access and access through a web-based interface, and intelligent and intuitive search mechanisms with a visual query designer. *IBM Tivoli* SIEM (Buecker et al., 2010) can provide a long-lasting and compact storage of information security events. The collected events are stored in a database as text objects containing information about incidents, management actions, correlation rules, etc. *Novell Sentinel Log Manager* stores all data in a compressed format (Novell, 2010). The components of data storage use a file-based storage and an indexing system. PostgreSQL is used for data management.

One of the alternative solutions on data representation in systems with complex data structures is the ontological approach (OWL, 2009). Using description logic, this approach can express much easier the complex relationships between entities. To represent an ontological meta-data we suggest using *RDFS* (RDF Schema) (RDF, 2004). RDF data model is a directed graph, which is based on elementary statements (triples). A *triple* is a short formal statement in the form of "subject-predicate-object". A *triple store* is a purpose-built database to store and retrieve RDF metadata (Triplestore, 2010). In addition to RDF and XML, OWL (Web Ontology Language) can be chosen to represent data. OWL is a language of Semantic Web, created to represent ontologies (OWL, 2009). OWL presents ontology in the form of documents, which can be stored and transmitted in a global network. *SWRL* (Semantic Web Rule Language) (SWRL, 2004) can be used to specify rules. SWRL is a proposal for a Semantic Web rule language, based on a combination of OWL

sublanguages with RuleML sublanguages (Parsia, 2005). SPARQL Protocol and RDF Query Language (SPARQL) is a query language to the data presented on the RDF model as well as the protocol for these requests and responses (SPARQL, 2008).

We also considered systems and approaches for *logical reasoning*, including Event Calculus (EC) based on first-order logic (through an example of the prototype we developed using CIFF and SICStus Prolog) (Kowalski et al., 1986); (Kakas et al., 2003) and Model checking based on linear temporal logic using SPIN (SPIN, 2012).

3 ONTOLOGICAL DATA MODEL

The use of ontology is necessary approach that enables to create a general model that can be flexibly and quickly all the necessary concepts in SIEM system in a particular area. Loose coupling of domain ontologies and a modular approach to development makes it easy to add, delete, and support individual ontologies. In addition, the components of ontologies may be dynamically combined according to requirements during a performance to meet specific application requirements. In addition, changes in the ontological data model require much less effort than the relational model. Therefore, it is particularly relevant in areas where it is needed to store different types of information that can quickly change. These include cyber security, as a whole, and SIEM systems in particular.

Teymourian et al. (Teymourian et al., 2009) consider the corporate semantic web technologies as a very promising direction to improve the efficiency of interaction between vendor services and their applications, due to the possibility of efficient usage of comprehensive enterprise-relevant knowledge. Li et al. (Li et al., 2010) propose an ontology-driven event processing framework as part of the middleware for smart spaces. The authors develop two key models that underlie their approach. As the basis to build the SIEM ontological model the data representation standards considered above are valuable to use.

There is a number of works which suggest the usage of these standards in security ontologies. For example, (Guo et al., 2009); (Parmelee, 2010); (Elahi et al., 2009) are devoted to building ontologies for SCAP protocol standards. (Heimbigner; 2011); (López de Vergara et al., 2004) consider the translation of CIM into the ontology representation. Within the task of designing the data models and the repository of SIEM systems for service infrastructures, we have developed an ontology to represent the data model for Attack Modeling and Security Evaluation Component (AMSEC). The SCAP protocol was taken as the basis to construct this model.

Figure 1 shows the ontology that describes the concepts (for vulnerabilities, software/hardware manufacturers and other concepts) and the hierarchy and relationships between these concepts. The ontology is made up of the data schema, which is called the TBox (Terminology box), and the data itself – ABox (Assertion box).

Description of the vulnerability is a certain sequence of hardware components connected by logical operators (AND, OR, NOT, AND, NOT OR). In the ontology such relationships are expressed as a set of axioms that allow bringing into the data model the possibility of logical reasoning. To specify the vulnerability in the relational data model is hard task. It is stored as a string which includes the entire list of vulnerabilities being parsed programmatically. This process takes a considerable amount of time and greatly increases the traffic to communicate with the repository. Applying the ontological approach allows solving the task of submitting such data much more efficient, reducing the sample size and speeding up the AMSEC functioning. We expanded the ontological model for vulnerabilities depicted in Figure 1 to specify the risk assessment, the countermeasures, and other concepts based on SCAP.

4 IMPLEMENTATION AND TESTING

Our proposals for implementation of the repository are, firstly, the recommendations on the choice of DBMS. Of course, traditional and popular relational DBMS (such as MySQL and PostgreSQL) together with XML-based DBMS can be used, but for the realization of an advanced ontology-based SIEM, which includes possibilities for logical reasoning and triplet stores are preferable.

In order to choose DBMS for repository, we investigated a set of relational, XML-based, and triplet stores. Regarding the choice of a relational DBMS, the most popular at the moment are Oracle, Microsoft SQL Server, Sybase, MySQL, PostgreSQL. A number of popular SIEM systems support these databases.

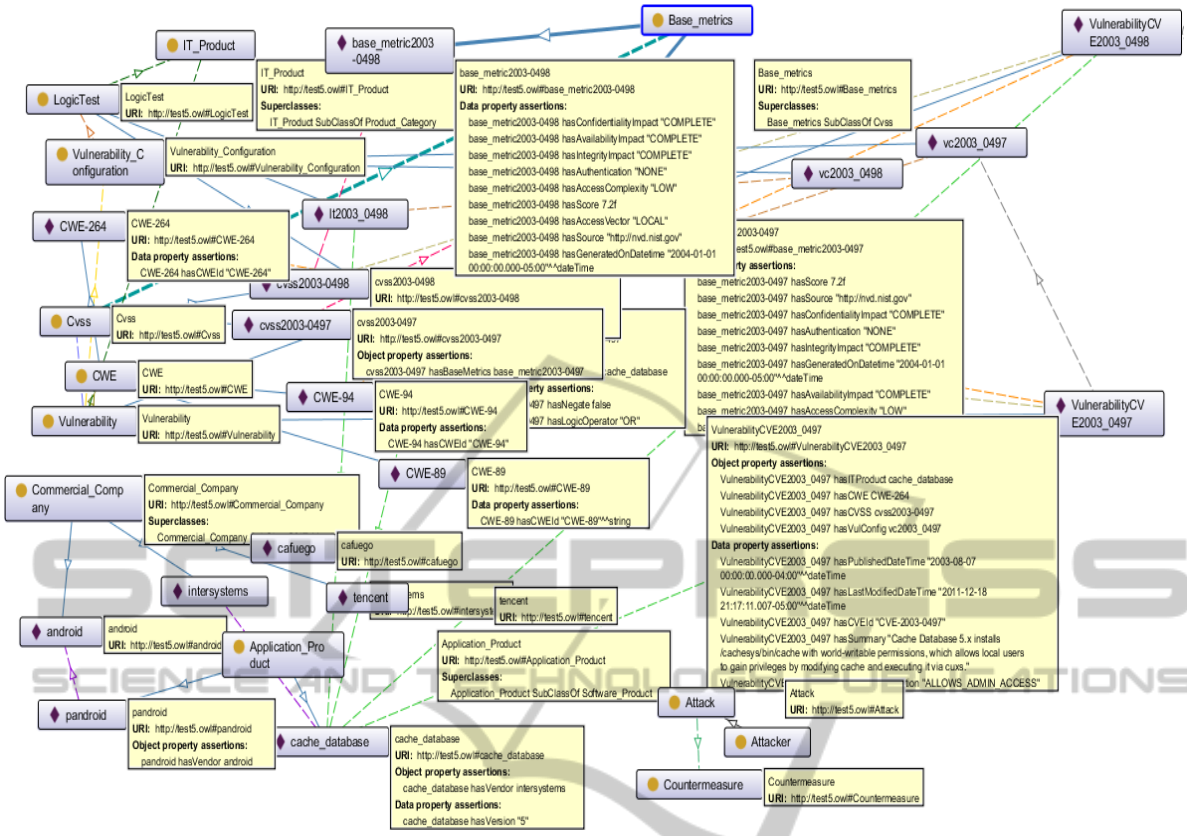

Figure 1: Ontological model for vulnerabilities.

Examples of XML-based data bases are Apache XIndice, BaseX, Sedna, Gemfire Enterprise, DOMSafeXML, eXist, MarkLogic Server, MonetDB/XQuery, OZONE, Xpriori XMS, etc.). Storage of triplets can be divided into two basic groups: implemented as standalone solutions (AllegroGraph, BigOWLIM and PelletDb), and parts of complex enterprise semantic system stores (Virtuoso, OpenAnzo and Semantics.Server).

According to the experience and the results of load tests, as the best practical solution for storage at the moment we proposed to use a hybrid approach that combines storage of triplets, relational and XML databases. This approach provides a balance in flexibility of data manipulation, the effective use of metadata and the acceptable processing speed.

The analysis of references shows that one of the best solution is Virtuoso (Virtuoso, 2012) by OpenLink Software Company. It combines the support of all three types of storages, has an open-source version that implements all necessary languages and protocols for data access and also supports a variety of necessary drivers.

SIEM modules access data through Web services, described by the WSDL standard. In its implementation, we have developed Web services for data access functions to CRUD (create, read, update, and delete). The WSDL is the link between the clients and the repository; it allows to correctly generate queries for accessing the repository. On the other hand, it is a specification for development of server-side Web service.

Implementation of the Web services was made in Java. All Web services are implemented as stateless, i.e. services do not share among themselves any variables and objects. This allows you to run the request from the client in a single thread on the application server. Thus, a single service can handle multiple threads of the same instances of classes. This allows speeding up the processing of requests and, if necessary, you can increase the number of servers. Any server can handle any request. To control this process use LoadBalancer which monitors the status of servers, receives requests from the client and forwards the request to the least-loaded server.

The repository testing was made according the integration repository with AMSEC of SIEM system. *Data repository updater* downloads the open databases of vulnerabilities, attacks, configuration,

weaknesses, platforms, and countermeasures from the external environment. *Specification generator* converts the information about network events, configuration and security policy, from other SIEM components or from users, into an internal representation. *Attack graph generator* builds attack graphs (or trees) by modeling sequences of malefactor's attack actions in the analyzed computer network using information about available attack actions of different types, services dependencies, network configuration and used security policy. *Security evaluator* generates combined objects of the attack graphs (routes, threats) and service dependencies, calculates the metrics of combined objects, evaluates the common security level, compares obtained results with requirements, finds "weak" places, generates recommendations on strengthening the security level. It performs stochastic imitation of multi-step attacks against the analyzed computer networks and determining the consequences with regard to various countermeasures and criteria defined by the decision maker (for example, security measures/tools effectiveness and efficiency against attacks, maintainability, reliability, operational costs, etc.). Security evaluator allows to select the solutions (validated events and alerts, possible future security events, countermeasures) needed for other MASSIF SIEM components. *Reports generator* shows vulnerabilities detected by AMSEC, represents "weak" places, generates recommendations on strengthening the security level and depicts other relevant security information.

5 CONCLUSIONS

The paper has examined the main issues of data model design and repository development for new generation SIEM systems. The main *innovations* we suggest are as follows: (1) *ontological approach* to provide the necessary flexibility of data representation in the repository and the possibility of more accurate and high-quality results of queering; (2) *hybrid approach* to implement the repository which allows to integrate relational databases, XML databases and stores of triplets; (3) advanced *repository architecture* implemented and tested with the data used for attack modeling in SIEM systems.

For this purpose, we analyzed the standards for processing information about events, the most common practical implementations of SIEM repositories, and ontology related research papers. We conducted a brief overview of various languages that can be used for data representation and manipulation. In addition, a comparative analysis of logical reasoning languages was fulfilled. This analysis allowed us to conclude that these languages can be used to implement the ontological approach.

The ontological vulnerability model was suggested. It is used in the repository for attack modeling and security evaluation. To implement the basic architecture of the repository, an approach based on Service-Oriented Architecture was chosen. We proposed to use a hybrid approach to storage repository that provides a balance in flexibility of data manipulation, the effective use of metadata and the acceptable processing speed. Our proposals for implementation of the repository are, firstly, the recommendations on the choice of DBMS. Of course, traditional and popular relational DBMS (such as MySQL and PostgreSQL) together with XML-based DBMS can be used, but for the realization of an advanced ontology-based SIEM, which includes possibilities of developed logical reasoning, the triplet stores are preferable.

Therefore, for full support of different information models being developed in the MASSIF, we suggest to use in the repository a hybrid approach. This approach combines the possibilities of relational, XML-based and triplet stores. As a practical solution, we propose to use the Universal Server Virtuoso. It combines the support of all three types of storages, has an open-source version that implements all necessary languages and protocols for data access and also supports a variety of necessary drivers.

The paper also considers the task of integrating the repository with other SIEM components through an example of developing and implementing Attack Modeling and Security Evaluation Component. Proposed ontological approach allows making AMSEC data more accurate, requires no further software processing, and thus improves the repository performance.

In further research we are planning to expand the proposed ontology, as well as to add to the repository different services that provide data security, verification of security properties and policies, etc. In addition, we are going to explore the issues of logical reasoning based on ontology repository, as well as the development of mechanisms for data visualization.

ACKNOWLEDGEMENTS

This research is being supported by grants of the

Russian Foundation of Basic Research (projects #10-01-00826 and #11-07-00435), the Program of fundamental research of the Department for Nanotechnologies and Informational Technologies of the Russian Academy of Sciences, the State contract #11.519.11.4008 and by the EU as part of the SecFutur and MASSIF projects.

REFERENCES

AccelOps, 2011. *AccelOps Security Information & Event Management (SIEM*). http://www.accelops.com/product/siem.php.

AlienVault, 2011. *AlienVault Unified SIEM System description*. AlienVault, Campbell, CA. 36 p.

Buecker, A., Amado, J., Druker, D., Lorenz C., Muehlenbrock, F., Tan, R., 2010. *IT Security Compliance Management Design Guide with IBM Tivoli Security Information and Event Manager*. IBM Redbooks.

CIM, 2011. *Common Information Model (CIM)*, DMTF. Website. http://dmtf.org/standards/ cim.

Elahi, G., Yu, E., Zannone, N., 2009. A Modeling Ontology for Integrating Vulnerabilities into Security Requirements Conceptual Foundations. In *ER'09 Proc. 28th International Conference on Conceptual Modeling*. Springer-Verlag Berlin, Heidelberg.

Guo, M, Wang, J, 2009. An Ontology-based Approach to Model Common Vulnerabilities and Exposures in Information Security. In *ASEE Southeast Section Conference.*

Heimbigner, 2011. D. DMTF - CIM to OWL: A Case Study in Ontology Conversion. http://www.docstoc.com/docs/23281194/DMTF---CIM-to-OWL-A-Case-Study-in-Ontology-Conversion.

Ingols, K., Chu, M., Lippmann, R., Webster, S., Boyer, S., 2009. Modeling modern network attacks and countermeasures using attack graphs. In *Proceedings of the 2009 Annual Computer Security Applications Conference (ACSAC '09)*, Washington, D.C., USA, IEEE Computer Society.

Kakas, A., Kowalski, R., Toni, F., 2003. Abductive Logic Programming. In *Journal of Logic and Computation,* V.2, No.6.

Kotenko, I., Stepashkin, M., 2006. Attack Graph based Evaluation of Network Security. In *Lecture Notes in Computer Science,* Vol. 4237, 2006.

Kowalski, R., Sergot, M., 1986. *A logic-based calculus of events*. New Generation Computing, V.4.

Li, Z., Chu, C.-H., Yao, W., Behr, R. A., 2010. Ontology-Driven Event Detection and Indexing in Smart Spaces. In *The 4th IEEE International Conference on Semantic Computing*, September 22-24, Carnegie Mellon University, Pittsburgh, PA, USA.

López de Vergara, J., Villagrá, V., Berrocal, J., 2004. Applying the Web Ontology Language to management information definitions. In *IEEE Communications Magazine*. Vol.42, pp.58-74.

Marco, D., Jennings, M., 2004. *Universal Meta Data Models*. Wiley.

MASSIF, 2011. Website. http://www.massif-project.eu.

Miller, D., Harris, S., Harper, A., VanDyke, S., Blask, C., 2011. *Security information and event management (SIEM) implementation*. McGraw-Hill Companies.

Novell, 2010. *Novell Sentinel Log Manager 1.0.0.5*. Installation Guide.

Ogle, D., Kreger, H., Salahshour, A., Cornpropst, J., Labadie, E., Chessell, M., Horn, B., Gerken, J., Schoech, J., Wamboldt, M., 2004. *Canonical Situation Data Format: The Common Base Event V1.0.1*. International Business Machines Corporation.

OWL, 2009. *OWL 2 Web Ontology Language Document Overview*. W3C Recommendation 27 October 2009. http://www.w3.org/TR/owl2-overview .

Parmelee, M, 2010. *Toward an Ontology Architecture for Cyber-Security Standards*. The MITRE Corporation.

Parsia, B., 2005. *Cautiously Approaching SWRL*. http://en.wikipedia.org/wiki/PDF.

Prelude, 2011. *Prelude Pro 1.0*. http://www.prelude-technologies.com/en/welcome/index.html

RDF, 2004. *RDF Vocabulary Description Language 1.0: RDF Schema*. W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-schema.

SCAP, 2011. The Security Content Automation Protocol (SCAP). Website. http://scap.nist.gov.

Shenk, J., 2009. *ArcSight Logger 4. Combat Cybercrime, Demonstrate Compliance and Streamline IT Operations*. A SANS Whitepaper. January 2009. http://www.arcsight.com/collateral/whitepapers/ArcSight_Combat_Cyber_Crime_with_Logger.pdf .

SPARQL, 2008. SPARQL Query Language for RDF. W3C Recommendation, 15 January 2008. http://www.w3.org/TR/rdf-sparql-query

SPIN, 2012. ON-THE-FLY, LTL MODEL CHECKING with SPIN. http://spinroot.com/spin/whatispin.html

Stevens, M, 2005. Security Information and Event Management (SIEM). In *The NEbraskaCERT Conference, August 9-11, 2005*. http://www.certconf.org/presentations/2005/files/WC4.pdf.

SWRL, 2004. *SWRL: A Semantic Web Rule Language Combining OWL and RuleML*. W3C Member Submission 21 May 2004. http://www.w3.org/Submission/SWRL/

Teymourian, K., Paschke, A., 2009. Towards Semantic Event Processing. In *Proceedings of the Third ACM International Conference on Distributed Event-Based Systems (DEBS '09)*. ACM. New York.

Triplestore, 2010. *Triple Store Evaluation Analysis Report*. Revelytix, Inc.

Vernooy-Gerritsen, M., 2009. *Emerging Standards for Enhanced Publications and Repository Technology*. Amsterdam University Press.

Virtuoso, 2012. http://virtuoso.openlinksw.com