УДК 004.056

И.В. Котенко, И.Б. Саенко, О.В. Полубелова

ПЕРСПЕКТИВНЫЕ СИСТЕМЫ ХРАНЕНИЯ ДАННЫХ ДЛЯ МОНИТОРИНГА И УПРАВЛЕНИЯ БЕЗОПАСНОСТЬЮ ИНФОРМАЦИИ

Котенко И.В., Саенко И.Б., Полубелова О.В. **Перспективные системы хранения данных для мониторинга и управления безопасностью информации.**
Аннотация. В статье приводится анализ наиболее известных и развитых в настоящее время систем хранения данных в части их использования для построения репозитория перспективных систем мониторинга и управления безопасностью информации (SIEM–систем). Анализу подвергаются реляционные СУБД, XML–базы данных и хранилища триплетов. Предложена и прокомментирована реляционная схема данных, интегрирующая аналитические модули SIEM–системы. Приведена классификация и характеристика известных средств построения и использования XML–баз данных. Среди хранилищ триплетов сделан выбор в пользу системы Virtuoso, обеспечивающей гибридный подход к построению репозитория в перспективных SIEM–системах, который был апробирован на решении задач моделирования атак и анализа защищенности.
Ключевые слова: безопасность информации, SIEM, система управления базами данных, XML–база данных, хранилище триплетов.

Kotenko I.V., Saenko I.B., Polubelova O.V. **Perspective data storage systems for security information monitoring and management.**
The summary. The paper analyzes the most well known and developed at present data storage systems that are used to build the repository for perspective security information monitoring and management systems (SIEM–systems). Relational DBMSs, XML-databases and stores are analyzed. The relational schema, that integrates analytical modules of SIEM system, is suggested and commented. The classification and characteristics of known tools of implementation and use of XML databases are given. Among triplet stores, the system Virtuoso is chosen. It provides a hybrid approach to implementation of the repository in perspective SIEM systems, which was probated for attack modeling and security analysis.
Keywords: information security, SIEM, data base management system, XML–data base, triplet store.

1. Введение. Наличие уязвимостей в компьютерных системах, разнообразие видов компьютерных атак, их непредсказуемый характер, территориальная и временная распределенность средств защиты сетевой инфраструктуры — все это приводит к тому, что в настоящее время для компьютерных сетей и систем все более важное значение приобретают технологии проактивной защиты информации, осуществляющих непрерывный мониторинг и управление безопасностью информации. В основе таких технологий лежит своевременный сбор данных о событиях безопасности, фиксируемых в записях журналов аудита компьютерной инфраструктуры, их хранение в специализированном хранилище и последующая обработка, включающая корреля-

цию, моделирование, выработку предупреждений и решений по противодействию атакам и восстановлению безопасности информации. Поэтому другим названием для системы, реализующей мониторинг и управление безопасностью информации, является термин *система управления информацией и событиями безопасности* (Security Information and Events Management, SIEM) [1, 2].

Центральным компонентом любой SIEM–системы является система хранения данных, или репозиторий [3]. В нем осуществляется хранение данных о событиях безопасности, представляющих собой информационную модель предметной области безопасности информации в компьютерной инфраструктуре, а также обработка запросов, поступающих от администраторов и аналитических модулей SIEM–системы. От того, какие решения приняты по выбору системы хранения и поддерживаемых ею моделей представления данных, зависят многие характеристики SIEM–системы: быстродействие, достоверность, непротиворечивость, способность осуществлять интеллектуальную обработку данных и другие.

В наиболее известных сегодня коммерческих SIEM–продуктах система хранения данных создается, как правило, на основе реляционной системы управления базами данных (СУБД). В то же время SIEM–системы нового поколения, проект создания которых под названием MASSIF (MAnagement Security and information in Service Infrastructures) выполняется в настоящее время в СПИИРАН совместно с рядом других научных организаций Евросоюза [4], требуют более эффективной организации данных в репозитории, чем та, которая предоставляется реляционной СУБД. В частности, одной из таких существенных возможностей является способность формировать модель данных в виде онтологии и осуществлять на ее основе логический вывод [5]. В качестве альтернативы реляционным СУБД либо как компоненты, дополняющие их возможности, в проекте MASSIF рассматриваются XML–базы данных и/или хранилища триплетов.

Целью настоящей работы является анализ применимости как традиционных реляционных СУБД, так и нетрадиционных, к числу которых относятся XML–базы данных и хранилища триплетов, к задачам, решаемым в современных и перспективных системах мониторинга и управления безопасностью информации, наиболее яркими представителями которых являются SIEM–системы,. Для реляционного подхода к построению репозитория предлагается и анализируется модель данных. Для XML–подхода проводится сравнительный анализ XML–СУБД. Для подхода, ориентированного на триплеты, выполняется ана-

лиз хранилищ триплетов, рассматриваются их архитектура и возможности, а также делается выбор наиболее приемлемого продукта для перспективных SIEM–систем.

2. Реляционный подход. Анализ решений по системе хранения данных, принятых в наиболее распространенных современных SEIM–системах, показывает, что реляционные СУБД остаются основными программно-инструментальными средствами построения репозитория. Рассмотрим наиболее характерные примеры реляционных решений по хранению данных в современных SIEM–системах.

SIEM–система AccelOps заменила систему MARS от Cisco [6]. Она предназначена для сбора записей журналов, генерируемых сетевыми устройствами и устройствами безопасности, разработанными как компанией Cisco, так и наиболее крупными сетевыми разработчиками. Записи журналов обрабатываются в режиме реального времени, классифицируются, коррелируются и хранятся для последующего online–поиска. В качестве СУБД используется реляционная PostgreSQL. Репозиторий реализован как online–система хранения, используемая для анализа записей журналов в режиме реального времени и для исторического анализа записей журналов.

Prelude Universal SIEM [7] является открытой SIEM–системой, предназначенной для сбора, нормализации, сортировки, агрегирования, корреляции и вывода на печать событий, связанных с обеспечением безопасности, независимо от торговой марки или лицензии, порождающей такие события. Она также формирует отчетность, используя исторические данные и мониторинг в реальном времени. Собранные данные хранятся в едином хранилище, которое поддерживается реляционными СУБД MySQL, PosgreSQL и SQLlite.

Система ArcSight, по мнению специалистов компании Gartner [8], считается наиболее развитой SIEM–системой. База данных ArcSight, реализуемая компонентом ArcSight DB, основана на СУБД Oracle 11g (11.2.0.2). В состав ArcSight DB входит программное обеспечение, позволяющее эффективно управлять данными, архивировать события и получать статус обработки данных и работы СУБД Oracle для самодиагностики системы [9].

SIEM–система IBM Tivoli обеспечивает долговременное и компактное хранение событий безопасности информации с помощью реляционной СУБД IBM DB2 [10]. Собранные события хранятся в базе данных как текстовые объекты, содержащие информацию об инцидентах, управленческих мерах, правилах корреляции и т.д.

SIEM–система Novell Sentinel хранит все данные в сжатом форма-

те [11]. Данные могут быть заархивированы локально или удаленно. Система может осуществлять полнотекстовый поиск по всем хранимым данным или поиск событий по определенному терму, например, по имени пользователя. Архитектура Novell Sentinel включает в себя коллекторы данных, хранилище данных, компоненты поиска и отчетности и пользовательские интерфейсы. Данные журналов поступают в хранилище данных от коллекторов данных. Компоненты хранилища данных используют систему индексации и хранения данных на основе файлов. Для управления данными используется PostgreSQL.

Как видно из проведенного обзора, существующие SIEM–системы, как правило, ориентированы на реляционный подход к моделированию данных и построению репозитория. В приведенных выше примерах, в частности, использовались следующие реляционные СУБД: открытые — PosgreSQL, MySQL и SQLlite; закрытые — Oracle и DB2. По всей видимости, реляционные СУБД в перспективных SIEM–системах также сохранят свои позиции. По этой причине нами была разработана реляционная модель данных для фрагмента предметной области перспективной SIEM–системы, которая охватывает описание уязвимостей, а также данные, используемые в двух аналитических компонентах — компоненте моделирования атак и анализа защищенности (Attack Modeling and Security Analysis Component, AMSEC) и прогностическом анализаторе безопасности (Prognostic Security Analyzer, PSA) [12, 13]. Схема данных этой модели приведена на рис. 1. Она использовалась в качестве примера для тестирования репозитория совместно с другими компонентами — AMSEC и PSA.

Приведенная схема данных может быть разделена на три части: модель данных AMSEC, модель данных PSA и интеграционный слой (Integration layer). Интеграционный слой используется обоими компонентами — AMSEC и PSA.

Схема данных модели AMSEC включает такие таблицы, как **network** (Сеть), **host** (Хост), **objectProperty** (Свойство объекта), **objectPropertyType** (Тип свойства объекта). Таблица **network** используется для хранения данных, описывающих анализируемую компьютерную сеть. Она содержит имя таблицы (*name*), ее описание (*description*), а также ссылку на таблицу **system** (Система), которая предназначена для описания системы как целого. Таблица **host** используется для спецификации хостов, составляющих сеть. Таблица **objectProperty** применяется для описания всех свойств хостов и сети. Таблица **objectPropertyType** содержит ссылки на таблицу **objectProperty**.

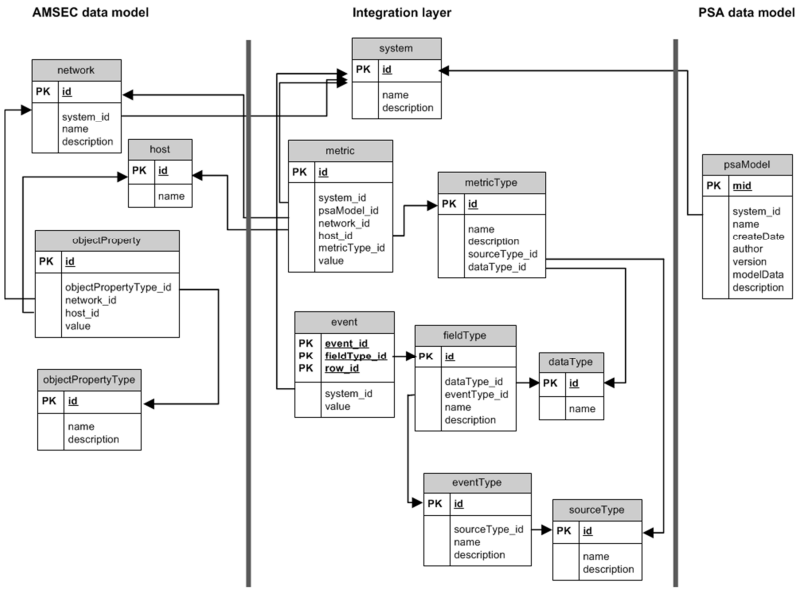

Рис. 1. Реляционная схема данных для SIEM–системы.

Схема данных модели PSA состоит из одной таблицы **psaModel**, которая используется для описания данных, используемых в PSA–анализе. Она содержит такие поля: *name* (имя), *description* (описание), *createDate* (дата создания), *author* (автор), *version* (версия), *modelData* (данные моделирования) и ссылку на таблицу **system**.

Кроме таблицы **system**, описанной выше, схема данных интеграционного слоя также содержит следующие таблицы: **metric** (Метрика), **metricType** (Тип метрики), **event** (Событие), **fieldType** (тип поля), **dataType** (Тип даты), **eventType** (Тип события), **sourceType** (Тип источника). Таблица **metric** содержит различные метрики, рассчитываемые с помощью PSA и AMSEC. Таблица **metricType** содержит имя и описание типа метрики, на который имеется ссылка в таблице **metric**. Остальные таблицы используются для хранения событий.

Приведенный пример демонстрирует, что реляционная схема данных может с успехом играть роль интегратора аналитических компонентов в SIEM–системе. Однако интеграция — не единственная задача, решаемая с помощью репозитория. В перспективных SEIM–системах должна быть реализована возможность использования внешних баз данных, описывающих уязвимости, угрозы, шаблоны атак и

прочее. В качестве примера можно привести внешнюю базу уязвимостей National Vulnerability Database (NVD) Национального института стандартов и технологии США (National Institute of Standards and Technology, NIST) [14]. База NVD является репозиторием, содержащим данные об уязвимостях, представленных в формате протокола Security Content Automation Protocol (SCAP) [15].

Декларируется, что база NVD обеспечивает возможность автоматизации управления уязвимостями и безопасностью. Помимо уязвимостей, база NVD содержит контрольные списки безопасности, описания связанных с безопасностью недостатков программного обеспечения и неправильных конфигураций, наименования продуктов и метрики безопасности. По состоянию на 27 января 2013 года в базе NVD, например, содержалось 54780 описания уязвимостей, и их количество постоянно возрастает.

Однако использование внешних баз, подобных NVD, при условии реляционной организации хранения данных может привести к существенному падению производительности SIEM–системы. Здесь следует выделить две причины.

Во-первых, необходимо периодически выполнять преобразование внешних данных, представленных в формате, предоставляемом SCAP, и их загрузку в реляционную базу данных. При этом следует учесть, что возможными форматами являются: текстовый, HTML, XML и XSD–схема (последний формат является языком описания структуры XML-документа). Следовательно, если бы репозиторий SIEM–системы мог хранить свои данные в формате XML или ему родственном, то такого рода вычислительных затрат можно было бы избежать или их существенно снизить.

Пример XML–описания уязвимости в формате SCAP приведен на рис. 2. Этот пример показывает, что уязвимость имеет место, когда хост имеет приложение «microsoft ie» и одну из следующих операционных систем: «microsoft vista sp2» или «microsoft vista sp2 x64», или «microsoft server 2008 sp2 x86», или «microsoft server 2008 sp2 x64», или «microsoft 7 x86», или «microsoft 7 sp1 x86», или «microsoft 7 x64», или «microsoft 7 sp1 x64», или «microsoft server 2008 r2 x64».

Второй причиной является тот факт, что описание уязвимости на языке SQL, являющемся базовым языком запросов для реляционных СУБД, как правило, является семантической конструкцией, в которой используются логические операторы OR и AND. Фрагмент такого описания уязвимости приведен на рис. 3.

```
<vuln:vulnerable-configuration id="http://nvd.nist.gov/">
  <cpe-lang:logical-test negate="false" operator="AND">
      <cpe-lang:logical-test negate="false" operator="OR">
          <cpe-lang:fact-ref name="cpe:/a:microsoft:ie:9"/>
      </cpe-lang:logical-test>
      <cpe-lang:logical-test negate="false" operator="OR">
          <cpe-lang:fact-ref name="cpe:/o:microsoft:windows_vista::sp2"/>
          <cpe-lang:fact-ref name="cpe:/o:microsoft:windows_vista::sp2:x64"/>
          <cpe-lang:fact-ref name="cpe:/o:microsoft:windows_server_2008::sp2:x86"/>
          <cpe-lang:fact-ref name="cpe:/o:microsoft:windows_server_2008::sp2:x64"/>
          <cpe-lang:fact-ref name="cpe:/o:microsoft:windows_7:::x86"/>
          <cpe-lang:fact-ref name="cpe:/o:microsoft:windows_7::sp1:x86"/>
          <cpe-lang:fact-ref name="cpe:/o:microsoft:windows_7:::x64"/>
          <cpe-lang:fact-ref name="cpe:/o:microsoft:windows_7::sp1:x64"/>
          <cpe-lang:fact-ref name="cpe:/o:microsoft:windows_server_2008:r2::x64"/>
          <cpe-lang:fact-ref
      name="cpe:/o:microsoft:windows_server_2008:r2:sp1:x64"/>
      </cpe-lang:logical-test>
  </cpe-lang:logical-test>
</vuln:vulnerable-configuration>
```

Рис. 2. Описание уязвимости в формате SCAP.

16	OR(cpe:/o:hp:apollo_domain_os:sr10.2,cpe:/o:hp:apollo_domain_os:sr10.3:beta)
17	OR(cpe:/o:sun:sunos:4.1,cpe:/o:sun:sunos:4.1.1)
18	OR(cpe:/o:sun:sunos:4.0.3,cpe:/o:sun:sunos:4.1,cpe:/o:sun:sunos:4.1.1)
19	OR(cpe:/o:sun:sunos:4.0.3,cpe:/o:sun:sunos:4.0.3c)
20	OR(cpe:/o:digital:ultrix:4.0,cpe:/o:digital:ultrix:4.1)
21	OR(cpe:/a:next:next:2.1)
22	OR(cpe:/o:att:svr4:4.0)
23	OR(cpe:/o:digital:ultrix:4.2)
24	OR(cpe:/a:ncsa:telnet)

Рис. 3. Описание уязвимости на языке SQL.

Каждая строка на рис. 3 соответствует операнду логического OR–выражения, требующего дополнительной процедурной обработки. Логично полагать, что такого рода описания, обрабатываемые при реляционном хранении, приводят к либо значительному структурному усложнению схемы данных, либо к усложнению SQL–запросов и процессов их обработки. Это связано с известной относительной статичностью реляционной модели и ее слабой приспособленностью к частым структурным изменениям в предметной области.

Как следствие, вновь для перспективной SIEM–системы возникает

необходимость либо полного перехода системы хранения с реляционной модели на альтернативную, в частности, на XML–модель, либо, как минимум, наличия дополнительной возможности непосредственного хранения и обработки данных в XML–формате. Для реализации таких возможностей в настоящее время имеются необходимые предпосылки, обусловленные наличием средств управления XML–базами данных. Рассмотрим их далее в рамках XML–подхода.

3. XML–подход. Формат XML изначально разрабатывался с целью обеспечения удобного и надежного обмена данными между разнородными системами хранения. В этой связи многие объекты компьютерной инфраструктуры, генерирующие события безопасности, имеют возможность представлять эти данные в XML–формате. Например, такой возможностью обладает Windows Server 2008, окно которого, описывающее свойства события в «режиме XML», показано на рис. 4. Это еще раз указывает на целесообразность наличия в перспективных SIEM–системах возможности хранения и обработки XML–данных.

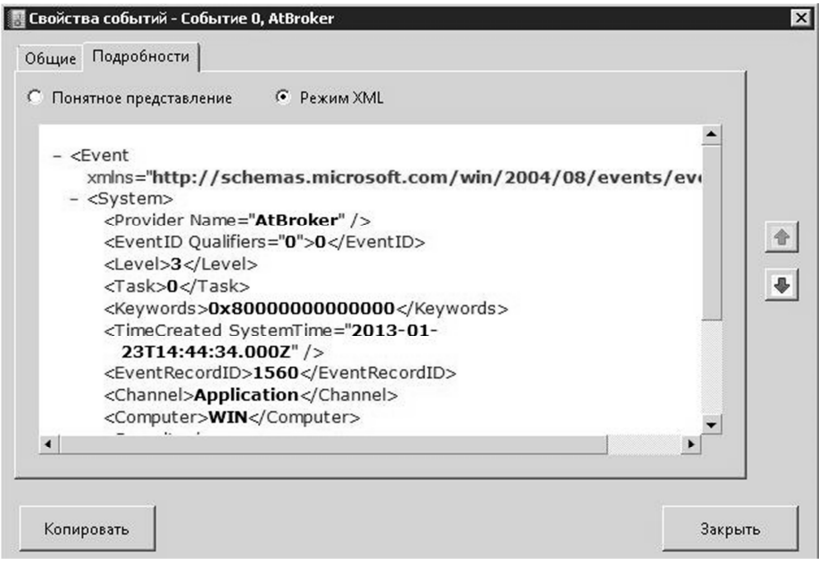

Рис. 4. Окно описания события безопасности в XML–формате.

Если еще десять лет назад был актуальным вопрос «Есть ли будущее у XML–СУБД?» [16], то обзор Р. Баррета по наиболее извест-

ным в мире средствам создания XML–баз данных в настоящее время включает 39 СУБД класса «естественных XML» (Native XML) [17] и 21 — класса «XML–поддерживающих» (XML–Enabled) [18].

Естественные XML–СУБД, в отличие от реляционных, определяют логическую модель для XML–документа, хранят и извлекают документы в соответствии с этой моделью. Данный класс СУБД использует XML–документ как единицу логического хранения данных.

Естественные XML–СУБД не требуют какой-либо модели физического хранения. Они могут быть встроены в реляционные, иерархические, объектно-ориентированные базы данных или использовать произвольный формат, например, индексные или сжатые файлы.

XML–поддерживающие СУБД содержат расширения, которые используются для преобразования данных между XML–документами и собственными структурами данных. Основное различие между XML–поддерживающими и естественными XML–СУБД заключается в том, что XML–поддерживающие базы данных используют структуры, определяемые схемами, которые должны быть отображены в XML–документе в ходе проектирования. Естественные XML–базы используют произвольные структуры, которые могут содержать произвольный XML–документ.

Еще одно отличие между этими двумя видами XML–СУБД содержится в используемом способе доступа к релевантным данным. Естественные XML–СУБД обращаются к данным через XML–ориентированные технологии, такие, как XQuery или XPath, и используют XML–ориентированные API, такие, как XQJ или XML:DB API.

XQuery (XML Query) — это язык запросов, разработанный для обработки данных в формате XML, который использует XML как свою модель данных. Он имеет статус официальной рекомендации (W3C Recommendation). В рамках стандарта SQL:2006 разработаны механизмы для встраивания XQuery–запросов прямо в SQL–запросы [19].

XPath (XML Path Language) — это язык запросов к элементам XML–документа, который является стандартом консорциума W3C. Он использует компактный синтаксис, отличный от принятого в XML–документах. В результате XPath–строка определяет путь к релевантным данным аналогично «пути к файлу» в «файловой системе» [20].

Естественные XML–СУБД можно классифицировать по следующим признакам: лицензия, типы данных, язык запроса и интерфейсы.

По признаку «лицензия» системы делятся на открытые, коммерческие, бесплатные и исследовательские. К открытым относятся: 4Suite, BaseX, Berkeley DB XML, DBDOM и другие. Коммерческие

системы: Dieselpoint, DOMSafeXML, EMC Documentum xDB и другие. Бесплатными являются: eXtc, Natix, Sedna XML DBMS и другие. К числу исследовательских относится система Lore.

По признаку «тип данных» деление ведется на собственные (proprietary) системы, являющиеся полностью естественными, и дополняющими, в которых XML не является основной моделью. К числу собственных относятся: BaseX, dbXML, EMC Documentum xDB, eXist, Natix, Sedna XML DBMS, Virtuozo и другие. Другими классами являются: объектно-ориентированные (4Suite, ozone, Sonic XML Server); реляционные (DBDOM); постреляционные (eXtc); основанные на файловой системе (DOMSafeXML, Extraway).

Язык XQuery используют системы: BaseX, Berkeley DB XML, EMC Documentum xDB, eXist, Sedna XML DBMS и другие.

Язык XPath используется следующими системами: 4Suite, dbXML, eXtc, Natix, Tamino, Virtuozo и другими.

В качестве интерфейсных средств различными системами используются: CORBA (4Suite, Xindice); SOAP (4Suite, EMC Documentum xDB, eXist, eXtc и другие); HTTP (4Suite, eXtc, MarkLogic Server, Tamino и другие); XQJ и XML:DB — практически все, имеющие тип данных «собственный».

Следует отметить, что неоднократно упоминавшаяся выше естественная открытая XML–СУБД Sedna XML DBMS является отечественной разработкой (Институт системного программирования РАН), которая имеет версии под Windows, Linux, Mac OS, FreeBSD и использует язык запросов XQuery. Ее особенностью является наличие специальных средств, с помощью которых Sedna может использоваться как «шлюз» с реляционным полем и выполнять SQL–запросы к реляционным данным через ODBC [21].

Среди XML–поддерживающих СУБД подавляющее большинство составляют реляционные СУБД. К их числу относятся: Access 2007(2010), DB2, FoxPro, Informix, MonetDB/SQL, MySQL, Oracle, PostgreSQL, SQL Server и Sybase ASE 15.0. Из них три системы являются открытыми, это — MonetDB/SQL, MySQL и PostgreSQL. Остальные СУБД являются коммерческими. Также имеются объектно-ориентированные системы (eXtremeDB, Matisse, Objectivity/DB, Orient ODBMS, Versant Object Database), вложенные реляционные (UniData и UniVerse), сетевые (RDM Embedded и RDM Server), постреляционная система Cache, многозначная система OpenInsight, ассоциативная система Sentences и некоторые другие.

Не умаляя значимость объектно–ориентированных систем, равно как и других нереляционных платформ баз данных, следует отметить, что, как видно из проведенного обзора, ведущие разработчики реляционной платформы — компании Oracle, IBM, Sybase и Microsoft — нарастили возможности своих СУБД способностью хранить информацию в XML–формате. СУБД, разрабатываемые этими компаниями, обеспечивают синтаксический разбор XML–информации в ходе ее импорта и ее сохранение в совокупности реляционных таблиц. Кроме того, они позволяют извлекать хранимые данные в XML–формате. В этой связи, перечисленные выше реляционные СУБД, являющиеся XML–поддерживающими, видятся как вполне приемлемые для перспективных SIEM–систем. Тем более, что они обладают высокой производительностью и способностью эффективно работать с терабайтными базами данных.

Однако естественные XML–СУБД способны поддерживать свободную форму представления информации. В случае их применения отпадает необходимость предварительного определения структуры XML–документов до их занесения в базу данных. В результате, естественные XML–СУБД позиционируются как наиболее подходящий инструментарий для взаимодействия с Web.

Это является достаточно весомым аргументом, чтобы рассматривать возможность применения естественных XML–СУБД в перспективных SIEM–системах, так как работа с внешними базами данных предполагает такое взаимодействие. Однако имеется еще один достаточно значимый довод в пользу данного заключения.

Так как естественные XML–СУБД способны работать с XML–документами произвольной структуры, с их помощью можно достаточно эффективно обрабатывать документы, созданные на основе различных XML–ориентированных языков, в частности, на основе языков Web Ontology Language (OWL) и Semantic Web Rule Language (SWRL). Язык OWL является языком Семантической Паутины (Semantic Web), созданным для представления онтологий [22]. Он группирует информацию в онтологию, которая представлена в виде XML–документов, Данные документы можно хранить в глобальной сети и передавать через нее. Эффективная обработка этих документов может быть выполнена средствами информации, скрытой внутри онтологии.

Язык SWRL (язык правил Семантической Паутины) является сочетанием подъязыков OWL DL и OWL Lite онтологического языка OWL с унарными/бинарными подъязыками Datalog RuleML языка правил разметки (Rule Markup Language) [23]. Тем самым он сохраняет

выразительную мощность языка OWL и дополняет ее возможностями языка RuleML.

Таким образом, оба рассмотренных выше языка — OWL и SWRL — являются XML–языками описания онтологий. Если система хранения данных основана на естественной XML–СУБД, следовательно, она способна хранить и обрабатывать XML–документы, сформированные на основе OWL и/или SWRL и, тем самым, способна хранить и обрабатывать онтологии предметных областей, в частности, предметной области управления событий безопасности.

Этот вывод является наиболее существенным обоснованием необходимости использования естественных XML–СУБД в перспективных SIEM–системах, так как возможность работы с онтологиями означает возможность логического вывода на данных о событиях безопасности [5]. В результате SIEM–системы обогащают свои возможности интеллектуальными сервисами защиты, а это становится их существенным достоинством в условиях кибер–противоборства [24, 25].

Однако, как бы ни были привлекательны естественные XML–СУБД в части их использования в SIEM–системах для хранения и обработки онтологий, в последнее время появился новый класс систем хранения данных, обладающий такой возможностью — хранилища триплетов. Приведем результаты их анализа в следующем разделе.

4. Подход, ориентированный на триплеты. Под *триплетом* понимается элементарное логическое утверждение вида «субъект — предикат — объект». Пример триплета покажем на утверждении «компьютер имеет ОС Windows 7». Здесь субъект — это «компьютер», предикат — это «имеет ОС», объект — это «Windows 7».

Триплеты являются основой построения предложенной консорциумом W3C модели представления данных Resource Description Framework (RDF) [26], предназначенной для записи утверждений о ресурсах различной природы в виде, пригодном для машинной обработки. RDF является частью концепции Семантической Паутины. Для обработки RDF–данных используются различные языки запросов. Языком запросов, рекомендуемым W3C, является SPARQL Protocol and RDF Query Language (SPARQL) [27].

Множество RDF–утверждений создают ориентированный граф, в котором вершинами являются субъекты и объекты, а ребра помечены предикатами.

Однако RDF–граф, взятый в отдельности, не раскрывает семантику описываемой предметной области. Для этой цели необходимы дополнительные средства. Одним их таких средств является RDF Schema

(RDFS) — специальный словарь для RDF, предназначенный для определения таксономий классов, свойств и т.д. [28].

RDFS имеет следующие достоинства:

- гибкость, т.е. любой пользователь может фиксировать произвольное утверждение о любом веб–ресурсе, используя RDF;

- открытость и расширяемость.

Однако следует отметить, что RDFS также свойственно отсутствие каких-либо гарантий целостности и последовательности RDF–описания.

Другим возможным средством интерпретации RDF–графов может являться упоминаемый выше язык описания онтологий OWL.

Таким образом, моделирование данных с помощью триплетов и реализация возможности их хранения и обработки в SIEM–системе является их достаточно привлекательной возможностью.

Системы хранения, ориентированные на работу с триплетами, получили название «хранилищ триплетов» (triplestores) [29]. Как и XML–СУБД, хранилища триплетов являются новым и интенсивно развивающимся направлением в области баз данных. Проведем обзор достигнутых ими на сегодня возможностей.

Хранилища триплетов можно разделить на две основные группы:

1) реализованные, как независимые решения (автономные);

2) являющиеся компонентом комплексной семантической системы хранения.

Примерами автономных решений являются AllegroGraph, BigOWLIM и PelletDb. Система AllegroGraph является наиболее популярным решением этого класса [30]. Она широко используется такими средствами, как TopBraid Composer (редактор онтологий), RacerPro (механизм вывода на языке OWL DL) и другими. Приведем ее характеристику.

Архитектура AllegroGraph (рис. 5) включает три уровня: уровень памяти, серверный и клиентский уровень. На уровне памяти находится RDF–хранилище AllegroGraph RDF Store. На серверном уровне находятся компоненты, обеспечивающие доступ различных платформ (Direct, HTTP, Sesame, SPARQL) к RDF–данным через общие серверные сервисы (Common Server Services). Клиентскую часть образуют средства создания интерфейсов (C#, Lisp, Java, Sesame, Jena, Clojure, Python, HTTP и т.д.).

Система AllegroGraph обладает хорошей информационной емкостью. Ее бесплатная версия способна хранить и обрабатывать до 50 миллионов триплетов.

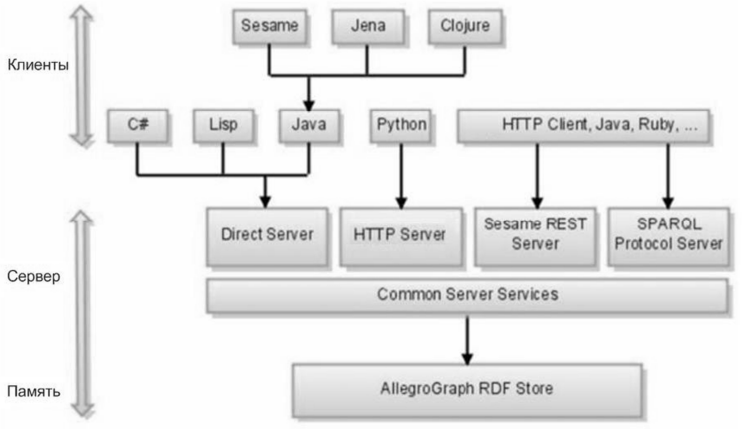

Рис. 5. Архитектура хранилища триплетов AllegroGraph.

Примерами второй группы хранилищ являются системы Virtuoso, OpenAnzo и Semantics.Server. Наиболее мощным из масштабируемых представителей этой группы является система Virtuoso [31]. Ее разработчиком является компания OpenLink, которая специализируется на разработке реляционных СУБД промежуточного программного обеспечения. Данная система имеет бесплатную версию.

Архитектура системы Virtuoso показана на рис. 6.

Как видно из этого рисунка, универсальный сервер Virtuoso состоит из модулей однородных хранилищ и модулей виртуальных баз данных.

Модули однородных хранилищ обеспечивают хранение данных в XML–формате, в реляционном виде (SQL–формате), в RDF–формате и в полнотекстовом виде.

Модели виртуальных баз данных играют роль СУБД и создают на основе данных, хранящихся в однородных хранилищах, полнофункциональные базы данных следующего назначения: XML–базы данных, хранилища триплетов в формате RDF, реляционные базы данных, веб–сервисы и полнотекстовые базы.

Сервер Virtuoso обеспечивает взаимодействие с компьютерными сетями (Internet / Intranet / Extranet) через большое количество платформ: ODBC, JDBC, OLEDB, .NET, HTTP, SOAP, SPARQL и другие.

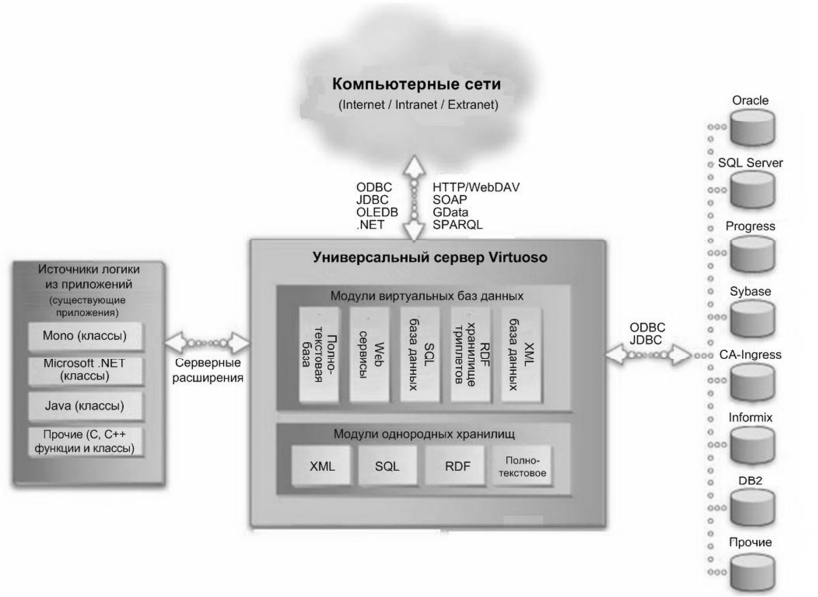

Рис. 6. Архитектура комплексного хранилища данных Virtuoso.

С помощью драйверов ODBC и JDBC универсальный сервер может взаимодействовать с достаточно большим количеством реляционных СУБД: Oracle, SQL Server, Progress, Sybase, CA–Ingress, Informix, DB2 и другими.

Наконец, через серверные расширения сервер может взаимодействовать с приложениями, разработанными на платформах Mono, .NET, Java, C, C++ и прочих, с целью импорта/экспорта логики, содержащейся в их классах и функциях.

Таким образом, Virtuoso с полным основанием может считаться комплексной системой хранения, так как, помимо RDF–данных, она обеспечивает хранение и интеграцию данных в других наиболее популярных форматах. В этом плане для перспективных SIEM–систем она выглядит более предпочтительной, чем система AllegroGraph.

Кроме того, система Virtuoso обладает достаточно высокой производительностью при работе с RDF–данными. Это подтверждается тестовыми оценками производительности различных систем хранения при обработке ими SPARQL–запросов, которые периодически проводятся в Берлинском университете Фрая (Berlin Freie University) на си-

стеме тестирования Berlin SPARQL Benchmark. Данная система является ориентиром для сравнения производительности различных систем хранения данных, предоставляющих точки доступа SPARQL. Она применяется для тестирования производительности RDF–хранилищ, систем отображения реляционных баз данных в RDF и SPARQL–приложений, ориентированных на другие типы данных [32].

Оценки показали, что производительность Virtuoso при выполнении запросов на выборке RDF–данных объемом 100M триплетов была выше, чем у остальных, в 1,5–2,5 раза (7352 запроса/час). При увеличении выборки в 2 раза производительность Virtuoso падала всего в 1,5 раза (4669 запросов/час), в то время как у большинства остальных систем — более чем в 2 раза.

Таким образом, и по своей производительности система Virtuoso является достаточно привлекательной. Учитывая тот факт, что одна из разработанных компанией OpenLink версий Virtuoso является бесплатной, данная система представляется наилучшим выбором системы хранения данных для перспективных SIEM–систем. Она позволяет реализовать *гибридный подход* к организации хранения данных [33], сочетающий как реляционные базы данных для отображения нормализованных данных о событиях, так и XML–документы, отображающие политики безопасности, шаблоны атак, инциденты и т.д., и хранилища триплетов, позволяющие работать с онтологиями. Это подтвердила основанная на Virtuoso реализация репозитория данных SIEM–системы, выполненная для сервисных инфраструктур, рассматриваемых в проекте MASSIF в качестве тестовых областей [34].

8. Заключение. Проведенный в настоящей работе анализ перспективных систем хранения данных для мониторинга и управления безопасностью информации показал, что класс реляционных СУБД не является единственным, заслуживающим внимания для реализации репозитория в SIEM–системах. В настоящее время имеется большое количество достаточно развитых естественных XML–СУБД, способных работать с XML–документами произвольного вида и тем самым расширять функциональность SIEM–систем в область интеллектуализации сервисов защиты информации. Кроме того, представляет значительный интерес использование хранилищ триплетов, класс которых в настоящее время также получил достаточное развитие. Хранение и обработка триплетов позволяет осуществлять достаточно развитую работу с онтологиями предметной области безопасности информации, что в еще большей степени оснащает SIEM–системы интеллектуальными средствами и сервисами защиты. Система Virtuoso, являющаяся

комплексной системой хранения разнородных данных, обеспечивает гибридный подход к построению репозитория и на наш взгляд является наилучшим выбором для перспективных SIEM–систем, так как она позволяет с высокой производительностью создавать и совместно использовать в SIEM–системах реляционные базы данных, XML–базы данных и базы (хранилища) триплетов.

Апробация предложенного гибридного подхода к построению репозитория SIEM–системы, ориентированной на функционирование в сервисных инфраструктурах для решения задач моделирования атак и анализа защищенности, подтвердила справедливость наших предложений.

Полученные результаты позволяют в дальнейших исследованиях перейти к разработке и тестированию модулей SIEM-системы, отвечающих за логический вывод и принятие решений на основе работы с онтологиями предметной области мониторинга и управления безопасностью информации.

Литература

1. *Котенко И.В., Саенко И.Б., Полубелова О.В., Чечулин А.А.* Применение технологии управления информацией и событиями безопасности для защиты информации в критически важных инфраструктурах // Труды СПИИРАН. СПб.: Наука, 2012. Вып. 1(20). С.27–56.
2. *Котенко И.В., Саенко И.Б., Полубелова О.В., Чечулин А.А.* Технологии управления информацией и событиями безопасности для защиты компьютерных сетей // Проблемы информационной безопасности. Компьютерные системы. 2012, № 2. С.57–68.
3. *Котенко И.В., Саенко И.Б.* SIEM–системы для управления информацией и событиями безопасности // Защита информации. Инсайд. 2012, № 5, С.54–65.
4. MASSIF FP7 Project. MAnagement of Security information and events in Service Infrastructures. http://www.massif-project.eu .
5. *Полубелова О.В., Котенко И. В., Саенко И.Б., Чечулин А.А.* Применение онтологий и логического вывода для управления информацией и событиями безопасности // Системы высокой доступности. 2012, № 2, т. 8. С.100–108.
6. 10 Reasons for Migrating from Cisco MARS to AccelOps. http://www.accelops.net/product/marsbeyond.php.
7. Prelude. Universal SIEM. http://www.prelude-ids.com/en/products/universal-siem/index.html.
8. *Nicolett M., Kavanagh K.M.* Magic Quadrant for Security Information and Event Management. Gartner. 2011.
9. Архитектура HP Arcsight. http://arcsight-russia.ru/products-hp-arcsight/architecture-hp-arcsight.
10. *Buecker A.; Amado J., Druker D., Lorenz C., Muehlenbrock F., Tan R.* IT Security Compliance Management Design Guide with IBM Tivoli Security Information and Event Manager. IBM Redbooks. 2010. 464 p.
11. Novell Sentinel Log Manager 1.0.0.5. Installation Guide. March 31, 2010. 38 p.

12. *Kotenko I., Chechulin A., Novikova E.* Attack Modelling and Security Evaluation for Security Information and Event Management // International Conference on Security and Cryptography (SECRYPT 2012). Rome, Italy. 24–27 July 2012. P.391–394.
13. *Kotenko I., Chechulin A.* Common Framework for Attack Modeling and Security Evaluation in SIEM Systems // 2012 IEEE International Conference on Internet of Things. Besançon, France, November 20–23, 2012. Los Alamitos, California. IEEE Computer Society. 2012. P. 94–101.
14. National Vulnerability Database Version 2.2. http://nvd.nist.gov/
15. The Security Content Automation Protocol (SCAP). http://scap.nist.gov/
16. *Дик Т.* Есть ли будущее у XML–СУБД? // PC Week, № 6(324), 2002. http://www.pcweek.ru/themes/detail.php?ID=60717.
17. Barret R. XML Database Products: Native XML Databases, 2010. http://www.rpbourret.com/xml/ProdsNative.htm .
18. Barret R. XML Database Products: XML–Enabled Databases, 2010. http://www.rpbourret.com/xml/ProdsXMLEnabled.htm.
19. XQuery 1.0: An XML Query Language (Second Edition). W3C Recommendation 14 December 2010. http://www.w3.org/TR/xquery/
20. XML Path Language (XPath) 3.0. W3C Candidate Recommendation 08 January 2013. http://www.w3.org/TR/xpath-30/
21. *Фомичев А., Гринев М., Кузнецов С.* СУБД Sedna: технические особенности и варианты использования // Открытые системы, № 08, 2004. http://www.osp.ru/os/2004/08/185085/
22. OWL 2 Web Ontology Language Document Overview (Second Edition). W3C Recommendation 11 December 2012. http://www.w3.org/TR/owl2-overview/
23. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission 21 May 2004. http://www.w3.org/Submission/SWRL/
24. *Котенко И.В., Саенко И.Б.* Построение системы интеллектуальных сервисов для защиты информации в условиях кибернетического противоборства // Труды СПИИРАН. СПб.: Наука, 2012. Вып. 3(22). С.84–100.
25. *Котенко И.В., Саенко И.Б., Юсупов Р.М.* Интеллектуальные сервисы защиты как инструмент кибернетического противоборства // Научно–технический сборник по проблемам информационного противоборства. М.: Совет Безопасности Российской Федерации. 2012.
26. Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation 10 February 2004. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
27. SPARQL Query Language for RDF. W3C Recommendation, 15 January 2008. http://www.w3.org/TR/rdf-sparql-query.
28. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-schema/
29. Triplestore. Wikipedia. http://en.wikipedia.org/wiki/Triplestore.
30. AllegroGraph 4.9. http://www.franz.com/agraph/allegrograph/
31. Virtuoso Universal Server. http://virtuoso.openlinksw.com/
32. BSBM V3 Results (February 2011). http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/results/V6/index.html
33. *Kotenko I., Polubelova O., Saenko I.* The Ontological Approach for SIEM Data Repository Implementation // 2012 IEEE International Conference on Internet of Things. Besançon, France, November 20–23, 2012. Los Alamitos, California. IEEE Computer Society. 2012. P.761–766.

34. *Kotenko I., Polubelova O., Saenko I.* Data Repository for Security Information and Event Management in service infrastructures // International Conference on Security and Cryptography (SECRYPT 2012). Rome, Italy, 24–27 July, 2012. P.308–313.

Котенко Игорь Витальевич — д-р техн. наук, проф., заведующий лабораторией проблем компьютерной безопасности, СПИИРАН. Область научных интересов: безопасность компьютерных сетей, в том числе управление политиками безопасности, разграничение доступа, аутентификация, анализ защищенности, обнаружение компьютерных атак, межсетевые экраны, защита от вирусов и сетевых червей, анализ и верификация протоколов безопасности и систем защиты информации, защита программного обеспечения от взлома и управление цифровыми правами, технологии моделирования и визуализации для противодействия кибер–терроризму. Число научных публикаций — более 450. ivkote@comsec.spb.ru, www.comsec.spb.ru; СПИИРАН, 14-я линия В.О., д.39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328–2642, факс +7(812)328–4450.

Kotenko Igor Vitalievich — Ph.D., Professor, Head of Laboratory of Computer Security Problems, SPIIRAS. Research interests: computer network security, including security policy management, access control, authentication, network security analysis, intrusion detection, firewalls, deception systems, malware protection, verification of security systems, digital right management, modeling, simulation and visualization technologies for counteraction to cyber terrorism; The number of publications — more 450. ivkote@comsec.spb.ru, www.comsec.spb.ru; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812) 328–2642, fax +7(812)328–4450.

Саенко Игорь Борисович — д-р техн. наук, проф.; ведущий научный сотрудник лаборатории проблем компьютерной безопасности СПИИРАН. Область научных интересов: автоматизированные информационные системы, информационная безопасность, обработка и передача данных по каналам связи, теория моделирования и математическая статистика, теория информации. Число научных публикаций — 250. ibsaen@comsec.spb.ru; СПИИРАН, 14-я линия В.О., 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328–2642, факс +7(812)328–4450.

Saenko Igor Borisovich — Ph.D., Professor; leading research scientist of Laboratory of Computer Security Problems, SPIIRAS. Research interests: automated information systems, information security, processing and transfer of data on data links, theory of modeling and mathematical statistics, information theory. The number of publications — 250. ibsaen@comsec.spb.ru; SPIIRAS, 14-th line, 39, St. Petersburg, 199178, Russia; office phone +7(812) 328–2642, fax +7(812)328–4450.

Полубелова Ольга Витальевна — научный сотрудник лаборатории проблем компьютерной безопасности СПИИРАН. Область научных интересов: безопасность компьютерных сетей, включая управление политиками безопасности, верификация протоколов безопасности и систем безопасности, использование методов проверки на модели для обнаружения и разрешения конфликтов в политиках; онтологии в информационной безопасности, дескрипционные логики, SIEM-системы. Число научных публикаций — 25. ovp@comsec.spb.ru, www.comsec.spb.ru; СПИИРАН, 14-я линия В.О., д.39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328–2642, факс +7(812)328–4450. Научный руководитель — Котенко И.В.

Polubelova Olga Vitalievna — researcher of Laboratory of Computer Security Problems, SPIIRAS. Research interests: computer network security, including policy management, verification of security protocols and security systems, model checking techniques for policy conflicts detection and resolution, ontology, description logic, SIEM systems. The number of publications — 25. ovp@comsec.spb.ru, www.comsec.spb.ru; SPIIRAS, 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328–2642, fax +7(812)328–4450. Scientific adviser — Kotenko I.V.

Поддержка исследований. В публикации представлены результаты исследований, поддержанные Министерством образования и науки Российской Федерации (государственный контракт 11.519.11.4008), грантами РФФИ, программой фундаментальных исследований ОНИТ РАН и проектами Седьмой рамочной программы Европейского Союза SecFutur и MASSIF.

Рекомендовано лабораторией криптологии, заведующий лабораторией Молдовян Н.А., д–р техн.наук, проф., заслуженный изобретатель РФ.
Статья поступила в редакцию 30.01.2013.

РЕФЕРАТ

Котенко И.В., Саенко И.Б., Полубелова О.В. **Перспективные системы хранения данных для мониторинга и управления безопасностью информации.**

В статье приводится анализ наиболее известных и развитых в настоящее время систем хранения данных в части их использования для построения репозитория перспективных SIEM–систем, являющихся наиболее яркими представителями систем мониторинга и управления безопасностью информации. Анализу подвергаются реляционные СУБД, XML–базы данных и хранилища триплетов. Предложена и прокомментирована реляционная схема данных, интегрирующая аналитические модули SIEM–системы. Приведена классификация и характеристика известных средств построения и использования XML–баз данных. Среди хранилищ триплетов сделан выбор в пользу системы Virtuoso, обеспечивающей гибридный подход к построению репозитория в перспективных SIEM–системах, который был апробирован на решении задач моделирования атак и анализа защищенности.

Проведенный обзор наиболее развитых существующих SIEM–систем (AccelOps, Prelude, ArcSight, IBM Tivoli, Novell Sentinel) показал, что реляционные СУБД в перспективных SIEM–системах сохранят свои позиции. Рассмотренная реляционная схема данных показала, что она может с успехом играть роль интегратора аналитических компонентов в перспективной SIEM–системе. Однако использование внешних баз при условии реляционной организации хранения данных может привести к существенному падению производительности SIEM–системы. Причины этого заключаются в необходимости периодического преобразования и загрузки в реляционную базу внешних данных, представленных в XML–формате, а также в трудоемкости обработки SQL–запросов, содержащих конструкции OR и AND.

Анализируемые средства построения XML–баз данных разбиты на два класса: естественные и XML–поддерживающие. Предложенная система классификации XML–СУБД использует следующие признаки: лицензия, типы данных, язык запроса и интерфейсы. Показано, что естественные XML–СУБД являются наиболее подходящими для взаимодействия с Web. Обоснована необходимость их использования для работы с XML–документами, сформированными на основе XML–языков работы с онтологиями (OWL и SWRL).

Дано понятие триплета и обоснована его связь с форматом RDF. Показана привлекательность моделирования данных с помощью триплетов и их хранения в SIEM–системе. Рассмотрены архитектуры двух основных классов хранилищ триплетов. На основе анализа тестовых данных показано, что комплексная семантическая система хранения данных Virtuoso обладает достаточно высокой производительностью при работе с RDF–данными. Обосновано, что система Virtuoso является наилучшим выбором системы хранения данных для перспективных SIEM–систем, реализующим гибридный подход к построению их репозитория.

SUMMARY

Kotenko I.V., Saenko I.B., Polubelova O.V. **Perspective data storage systems for security information monitoring and management.**

The paper analyzes the most well known and developed at present data storage systems that are used to build the repository for perspective security information monitoring and management systems (SIEM–systems). Relational DBMSs, XML-databases and stores are analyzed. The relational schema, that integrates analytical modules of SIEM system, is suggested and commented. The classification and characteristics of known tools of implementation and use of XML databases are given. Among triplet stores, the system Virtuoso is chosen. It provides a hybrid approach to implementation of the repository in perspective SIEM systems, which was probated for attack modeling and security analysis.

A review of the most advanced existing SIEM-systems (AccelOps, Prelude, ArcSight, IBM Tivoli, Novell Sentinel) showed that the relational DBMSs in perspective SIEM-systems will retain their positions. The considered relational data schema has shown that it can successfully play the role of an integrator of analytical components in perspective SIEM-system. However, the use of external databases under conditions of the relational organization of the data storage can lead to significant performance degradation of the SIEM system. The reasons for this lie in the need for the periodic conversion and loading in the relational database the data represented in XML-format, as well as in the complexity of processing SQL-queries that contain structures OR and AND.

The most popular and developed tools for XML databases are divided into two classes: natural and XML–enabled. A proposed classification system for XML databases used the following signs: license, data types, query language and interfaces. The paper shows that natural XML-DBMS are being positioned as the most suitable tools to interact with the Web. We emphasize the necessity of using natural XML databases to work with XML documents, which are created on the basis of XML-based ontology language (OWL and SWRL).

The notion of the triplet is given and its relationship to the RDF format is justified. The paper shows that the data representation with triplets and the realization of triplet storage and processing in SIEM-system are a sufficiently attractive option. The architecture and characteristics of triplet stores have discussed for two major classes: implemented as independent decisions (AllegroGraph system) and being as a part of a complex semantic storage system (Virtuoso system). Based on the analysis of test data the paper shows that the Virtuoso has high enough performance when it works with RDF-data. The paper substantiates that the Virtuoso is the best choice of the storage system for perspective SIEM systems that use the hybrid approach to implement their repositories.