S. V. Zapechnikov, N. G. Miloslavskaya, M. U. Senatorov, A. I. Tolstoy

Information Security Maintenance Issues for Big Data

Keywords: big data, information security, secure infrastructure, data visualization

The main big data technologies are briefly listed. The need to protect such data, particularly those relating to information security (IS) maintenance of an enterprise's IT infrastructure, and their processing processes is shown. A worldwide experience of addressing big data IS issues is briefly summarized. A big data protection problem statement is formulated. An infrastructure for big data information security maintenance is offered. New applications areas for big data IT after addressing IS maintenance issues for them are listed in conclusion.

С. В. Запечников, Н. Г. Милославская, М. Ю. Сенаторов, А. И. Толстой

ПРОБЛЕМЫ ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ БОЛЬШИХ ДАННЫХ

Введение

Современные корпорации разного размера, подчинения и сферы деятельности получают огромное количество данных о текущем состоянии их ИТ-инфраструктуры (ИТИ) и, на первый взгляд, никак не связанных (разрозненных) событиях, происходящих в ней. Эти данные требуется корректно и оперативно обработать для выявления инцидентов информационной безопасности (ИБ) и выделения областей ИТИ, наиболее подверженных высокому риску, для его оперативного устранения. Данные формируются из информации, рассматриваемой в определенном контексте, поступающей от отдельных контроллеров доменов, прокси-серверов, DNS-серверов, средств защиты информации (СЗИ), описывающей текущую конфигурацию сетевых устройств, характеристики сетевого трафика, работу приложений и сетевых сервисов, активность и конкретные действия отдельных конечных пользователей, а также содержащей почтовую переписку, веб-контент, оцифрованный звук и видео, данные бизнес-процессов, внутренние документы и аналитические данные корпорации за многие годы ее существования.

Объемы и неоднородность подлежащих дальнейшему тщательному мониторингу и анализу данных и связанной с ними активности очень велики. Вопрос их структурированного, консолидированного и визуального представления для принятия своевременных и обоснованных решений в области управления ИБ всеми активами ИТИ корпорации встает очень остро. Постоянно

возрастающие объемы данных о событиях ИБ, активах ИТИ, их уязвимостях, пользователях, угрозах ИБ и сопутствующей информации, а также необходимость более быстрого получения систематизированной и определенным образом проанализированной «сырой» (англ. raw) разнородной информации для более быстрого понимания текущей ситуации в обеспечении ИБ ИТИ породило известную проблему больших данных.

Совокупность сложившихся в этой области технологий получила название Big Data. Критериями, определяющими отличие больших данных от ставших традиционными ИТ, являются «три V»: 1) volume — сверхбольшой объем данных, 2) velocity — очень высокая скорость передачи данных, 3) variety — слабая структурированность данных, которая понимается, прежде всего, как нерегулярность структуры данных и трудность извлечения однородных данных из потока и выявления корреляции. Позднее к ним добавились и другие критерии: veracity (достоверность), variability (изменчивость), value (ценность), visibility (обозримость).

К настоящему времени общепринятой терминологии в области обработки больших данных еще не сложилось. Однако можно предложить следующую интерпретацию понятия «большие данные». Это массивы данных такого объема и структуры, которые превышают возможности традиционных программных инструментов (баз данных, ПО) по сбору, хранению и обработке данных за приемлемое время и тем более превышают возможности их восприятия человеком. При этом данные могут быть структурированными, слабоструктурированными и неструктурированными, что не позволяет эффективно управлять ими и обрабатывать традиционным образом.

Общепризнанно, что технологии поддержки распределенных вычислительных сред, включая большие данные, разрабатывались без учета требований обеспечения ИБ для них. Поэтому этот вопрос особо актуален, но еще далек от полного разрешения.

Проблемы обеспечения ИБ для ИТ больших данных, относящихся к ИБ ИТИ, и являются предметом рассмотрения в данной статье. Ее структура такова. В разделе 1 кратко перечисляются основные технологии больших данных. Раздел 2 обосновывает необходимость защиты таких данных, особенно относящихся к обеспечению ИБ ИТИ корпорации, и процессов их обработки. Раздел 3 кратко обобщает опыт разработки вопросов защиты для технологий больших данных. В разделе 4 формулируется постановка задачи защиты больших данных. В разделе 5 предлагается инфраструктура обеспечения ИБ больших данных в указанной предметной области. Новые области применения ИТ больших данных после решения вопросов обеспечения ИБ для них перечислены в заключении.

1. Технологии больших данных

Можно выделить два типа обработки больших данных [1]:

1) пакетная обработка (англ. batch processing) в псевдореальном, или «мягком» реальном, времени, когда обработке подвергаются данные, уже сохраненные в энергонезависимой памяти, а вероятностно-временные характеристики процесса преобразования данных, в основном, определяются требованиями прикладных задач;

2) потоковая обработка (англ. stream processing) в «жестком» реальном времени, когда обработке подвергаются данные, находящиеся в оперативной памяти, без сохранения на энергонезависимых носителях, а вероятностно-временные характеристики процесса преобразования данных, в основном, определяются темпом поступления данных, так как появление очередей на обрабатывающих узлах приводит к безвозвратной утрате данных.

Большие данные следует понимать прежде всего как ИТ, направленные на обработку сверхбольших массивов слабоструктурированных данных в режиме реального времени.

ИТ больших данных в корне отличаются от традиционных ИТ тем, что они становятся «дата-центрированными», или «управляемыми данными» (англ. data-centric, data-driven). Если

в традиционных ИТ в центр процесса обработки данных ставилось обрабатывающее устройство или среда (компьютер, кластер, облачная среда), которые обрабатывали запросы (заявки и пр.), то большие данные рассматриваются прежде всего как непрерывно «текущая» субстанция, механизмы обработки которой должны быть встроены в потоки. При этом скорость приема вновь поступающих данных в обработку и скорость выдачи результатов должны быть не ниже скорости потока, так как в противном случае это приводило бы к бесконечному росту очередей либо бесполезному хранению бесконечно увеличивающегося объема необработанных данных.

Теоретической основой ИТ больших данных является раздел информатики (computing), получивший название «наука о данных» (англ. data science), включающий в себя следующее [2]:

1) разработка методологии распределенных файловых систем и преобразования массивов данных путем использования методологии отображения и свертки (MapReduce) в качестве инструментария создания процедур параллельной и распределенной обработки сверхбольших массивов данных;

2) поиск по сходству (англ. similarity search), включая ключевые технологии минхэширования (англ. minhashing) (поиск пересечений в подмножествах массива) и локально-чувствительного хэширования;

3) обработка потоковых данных и специализированные алгоритмы для быстроприбывающих данных, которые должны быть либо немедленно обработаны, либо безвозвратно потеряны;

4) технологии поиска по сверхбольшим массивам данных и ранжирования результатов поиска (типа Google's PageRank);

5) обнаружение часто повторяющихся подмножеств данных (англ. frequency-itemset data mining), включая ассоциативный поиск, метод «рыночных корзин» и их улучшения;

6) алгоритмы кластеризации сверхбольших массивов высокой размерности;

7) проблемы веб-приложений: поиск адресатов для эффективной рассылки информации и прогнозирование предпочтений пользователей на основе изучения их активности в Интернете;

8) алгоритмы анализа и извлечения структуры очень больших графов, в особенности графов социальных сетей;

9) методы получения качественных и количественных характеристик больших массивов данных путем понижения (редукции) размерности, включая декомпозицию к единственной величине, латентную (скрытую) семантическую индексацию и различные виды корреляций;

10) алгоритмы машинного обучения, которые могут быть применены к большим данным.

2. Необходимость обеспечения информационной безопасности больших данных

Бывший руководитель группы сервисов защиты корпорации Barclays, в настоящее время работающий в компании Splunk, С. Гейли (S. Gailey) сказал: «Безопасность в настоящее время является проблемой больших данных, так как данные, которые имеют контекст безопасности, огромны. Это не просто набор средств безопасности, производящих данные, это вся ваша организация. Если вы собираетесь игнорировать некоторые из этих данных или если вы не можете проанализировать их, то вы не защищаетесь должным образом. Каждая мелочь, которую вы пропустите или игнорируете, важна для вашей компании» [3].

Актуальность проблемы обеспечения ИБ больших данных определяется противоречием между возрастающими потребностями в обработке таких данных, возникающими в различных сферах человеческой деятельности, и недостаточными возможностями гарантировать конфиденциальность, целостность, доступность обрабатываемых данных и, в частности, персональных данных, а также обеспечение безопасности средств обработки данных и компонентов инфраструктуры, включая системное и прикладное ПО центров обработки больших данных.

В некоторых публикациях выражено мнение, что перечисленные задачи можно решить, применяя хорошо известные подходы и привычные средства защиты. По нашему мнению,

радикальная перестройка ИТ применительно к задачам обработки больших данных вызывает необходимость выработки принципиально иных подходов к обеспечению ИБ с соответствующей методологической базой, в настоящее время находящейся на этапе становления. Обоснованность нашего мнения связана с тем, что традиционные технологии обеспечения ИБ строятся на концепции защиты единой физической сущности (типа сервера или базы данных), а не распределенной, крупномасштабной вычислительной среды больших данных.

Проблемы обеспечения ИБ больших данных, по мнению ряда авторитетных источников (в частности, согласно документам Рабочей группы по большим данным Альянса облачной безопасности — англ. Cloud Security Alliance), можно подразделить на четыре категории:

- обеспечение безопасности инфраструктуры больших данных;
- обеспечение логической целостности и мониторинга безопасности в режиме реального времени;
- обеспечение тайны личной жизни, или приватности (англ. privacy);
- рациональное управление данными.

Сам объект обработки в рамках технологий больших данных — в нашем случае данные об ИБ — является «чувствительным», поскольку содержит важную информацию о состоянии ИБ корпорации. Большие объемы информации, связанной с безопасностью, должны быть собраны, проанализированы, классифицированы, нормализованы и скоррелированы для определения рисков ИБ. После обработки в конкретном контексте эти данные приобретают еще больший смысл, становятся значимой информацией, которую необходимо скрывать от злоумышленников. Именно эти данные и являются самым важным объектом защиты в технологиях больших данных. Чем больше значимой информации расположено и перемещается между большим количеством мест, чем когда-либо прежде, тем сильнее увеличивается количество точек и способов взломать ее, заканчивающихся кражей, мошенничеством, потерей репутации и т. д. Поэтому так необходима защита как самого процесса обработки больших данных, так и всех его входных данных и результатов.

Ключевым моментом обоснования необходимости обеспечения ИБ технологий больших данных, как и любой другой технологии, является риск-ориентированный подход. Но полноценный анализ рисков для больших данных — это предмет отдельного большого исследования, который не может быть изложен в рамках одной статьи. Поэтому выделим лишь некоторые риски, связанные с технологиями больших данных [4]:

- как любая новая технология, «большие данные» не очень хорошо поняты компаниями и вводят новые уязвимости;
- ее реализации, как правило, включают в себя открытый исходный код, часто имеющий ненайденные (до некоторого времени) лазейки и учетные данные по умолчанию (собственное ПО также имеет лазейки и учетные данные по умолчанию, но их труднее найти без открытого исходного кода); если злоумышленник найдет некоторые уязвимости в открытом исходном коде раньше, чем другие, то он может использовать его для получения прибыли, пока они не будут устранены;
- нерассмотренные атаки на узлы кластера и неадекватно защищенные серверы;
- недостаточный контроль за аутентификацией пользователя и доступом к данным из различных мест;
- невыполненные нормативные требования, с проблемами доступа к логам и данным аудита;
- вредоносные входные данные и неадекватная проверка данных и т. д.

Конечно, этот список должен быть значительно расширен, например, виртуальным и распределенным характером самих больших данных.

3. Исследования, ведущиеся в области обеспечения информационной безопасности больших данных

В мире ведутся интенсивные исследования проблемы обеспечения ИБ технологий больших данных в целом.

В 2012 г. правительство США анонсировало стратегическую «Инициативу больших данных», в которую вовлечено шесть федеральных министерств и ведомств страны. Основным исполнителем работ является Университет Беркли (Калифорния). В Массачусетском технологическом институте создан специальный Научно-технологический центр по большим данным (Science and Technology Center for Big Data) в структуре Лаборатории компьютерных наук и искусственного интеллекта.

Исследования в области методологии криптографической защиты больших данных ведутся группой по прикладной криптографии в составе Лаборатории безопасности Департамента компьютерных наук Стэнфордского университета (США).

Под эгидой Европейской Комиссии осуществляется проект «Big Data Public Private Forum» в рамках стратегической инициативы Horizon 2020. Научные работы по рассматриваемой тематике публикуются интернациональным научным коллективом, включающим ученых из ФРГ, Нидерландов, Франции, Швейцарии, Дании.

Проекты в области обеспечения ИБ технологий больших данных активно осуществляются ведущими компаниями — производителями продуктов и систем ИТ. В частности, результаты исследований компании IBM уже реализованы в серии продуктов InfoSphere Guardium для защиты структурированных и неструктурированных данных как в режиме реального времени, так и в «отложенном» (offline) режиме.

Ввиду широты и многогранности проблемы не все отдельные задачи получают одинаковое внимание. Основные направления исследований в указанной области в настоящее время таковы.

1. Разработка методов обеспечения конфиденциальности информации, обрабатываемой в недоверенных средах, путем выполнения операций «под шифром». Достигнутый в настоящее время уровень решения этой проблемы теоретически позволяет выполнять «под шифром» операции, которые описываются произвольными булевыми функциями (с определенными ограничениями) [5], однако сложность соответствующих алгоритмов и используемый в них математический аппарат пока не обеспечивают (даже близко) уровень производительности, пригодный для их практического применения, а тем более для применения в системах реального времени.

2. Разработка механизмов контроля доступа к большим данным. Традиционные методы контроля доступа, основанные на дискреционной, мандатной и ролевой моделях, оказываются неприменимы при обработке больших данных. Ведущим направлением исследований здесь является создание новых моделей контроля доступа на основе атрибутной модели [6].

3. Разработка методов и средств дистанционного контроля целостности массивов данных, обрабатываемых во внешней по отношению к владельцу данных и потому недоверенной среде. Вследствие трудности реализации процедур полного контроля ведущее значение здесь имеют методы вероятностного контроля целостности. Однако все известные методы рассчитаны только на контроль статических массивов. Получают развитие так называемые локально-чувствительные функции хэширования (англ. locality-sensitive hashing) и локально-декодируемые коды [7], позволяющие контролировать участки большого массива, но их применение к быстро изменяющимся массивам данных требует проведения дополнительных исследований.

4. Создание методов доказательной регистрации событий и отслеживания происхождения («трассировка») быстро изменяющихся массивов данных (англ. data provenance) [8].

5. Создание методов обработки данных «по доверенности» (как правило, имеется в виду ограниченный набор операций) и верификации результатов вычислений. Самым значительным достижением в этой области на сегодняшний день является компилятор программ, написанных на языке C, с открытым исходным кодом, созданный научным коллективом компании Microsoft и ряда университетов [9].

6. Создание методов защищенного сбора, обработки и хранения свидетельств событий ИБ, представленных в электронной форме (англ. digital evidence) при обработке больших данных в целях исключения их фальсификации и принятия решений по управлению ИБ [10].

7. Исследование проблемы безопасных двусторонних и многосторонних вычислений при различных начальных условиях и исходных установках (безопасность понимается в смысле конфиденциальности и аутентичности не только начальных и конечных данных, но и всех промежуточных результатов выполняемых вычислений) [11].

Большинство методов, разработанных в ходе исследования перечисленных выше проблем, имели целью проверку принципиальной возможности реализовать заявленную функциональность. Имеющиеся в настоящее время решения применимы преимущественно к обработке данных в псевдореальном времени («мягком» реальном времени).

Предложенные на сегодняшний день методы и алгоритмы, как правило, предназначены для обеспечения либо некоторых аспектов ИБ для постоянно хранимых массивов данных или их отдельных элементов, либо безопасности средств обработки больших данных и инфраструктуры, на которой осуществляется их обработка. Такие методы ориентированы обычно на применение при пакетной обработке в рамках методологии MapReduce.

В то же время соответствующие методы и алгоритмы, ориентированные на потоковую обработку, в данный момент практически отсутствуют. Исследования в этой области только начинаются научными группами в ведущих университетах и научно-исследовательских подразделениях фирм — производителей аппаратного и программного обеспечения.

4. Постановка задачи защиты больших данных

Отсутствие на сегодняшний день системного подхода к обеспечению ИБ процесса обработки больших данных в режиме реального времени сдерживает применение этих ИТ.

Задача обеспечения ИБ больших данных может быть корректно сформулирована следующим образом: создать комплекс моделей и методов, позволяющих обеспечить защищенность процесса обработки больших данных в режиме «жесткого» реального времени на обрабатывающих узлах и в центрах обработки данных (ЦОД).

В качестве начального условия решаемой задачи предполагается, что на вход узла поступает с высокой скоростью поток слабоструктурированных данных (в том числе, возможно, образующихся путем смешения нескольких отдельных потоков). ЦОД осуществляют сплошную обработку входящего потока со скоростью не ниже, чем скорость входящего потока, без потери какой-либо части потока. Технология обработки потока может представлять собой произвольное количество отдельных элементарных операций, на вход каждой из которых поступает либо исходных поток, либо поток, являющийся результатом какой-либо уже выполненной элементарной операции. На выходе каждая элементарная операция порождает поток, который либо поступает на вход другой элементарной операции, либо является выходящим потоком ЦОД.

Таким образом, функции ЦОД описываются графом, вершинами которого являются элементарные операции, а ребрами — потоки между ними. Конкретное содержание выполненных элементарных операций и конфигурация потоков между ними определяются прикладными задачами, которые выполняет ЦОД. Интенсивность выходных потоков не обязательно совпадает с интенсивностью входных потоков.

В этих условиях требуется, во-первых, обеспечить (возможно, в рамках задаваемых по отдельным показателям количественных требований) защищенность проходящего через ЦОД информационного потока, включая как основные традиционные аспекты ИБ (конфиденциальность, целостность и аутентичность данных), так и специфические новые требования, возникающие в связи с характеристиками потока данных (большой объем, слабая структурированность, высокая

интенсивность) и процедур его обработки. К таким требованиям могут, в частности, относиться требования контроля доступа к отдельным элементам или структурам данных, приходящим в потоке; отслеживание персональных данных в потоке или взаимосвязей между анонимными данными, позволяющих достоверно установить лиц, с которым они связаны.

Вторая группа требований относится к алгоритмической составляющей процесса обработки: требуется обеспечить доверие к компонентам ЦОД, реализующим отдельные элементарные операции, а если обеспечение достаточного уровня доверия к ним невозможно, то верификацию результатов выполнения операции и коррекцию случайных ошибок либо преднамеренно внесенных искажений.

Наконец, третья группа требований относится к обеспечению безопасности инфраструктуры ЦОД. Самый распространенный в настоящее время подход к обработке больших данных, независимо от актуальности требования обработки в режиме реального времени, — это создание кластера вычислительных средств и программная реализация массово-параллельной обработки (англ. massively parallel processing) элементов сверхбольшого массива [12]. При отсутствии у потребителя собственной ИТИ он может пользоваться услугами провайдера облачных вычислений. В этом случае облачные вычисления предоставляются по модели «инфраструктура как сервис» (IaaS), так как другие модели («платформа как сервис» и «ПО как сервис»), очевидно, нерациональны или вообще невозможны при наличии требования режима реального времени.

Решаемая задача в такой постановке является одной из ключевых задач, ведущих к решению на удовлетворительном уровне проблем обеспечения ИБ технологий больших данных.

Научная новизна поставленной задачи определяется следующими факторами:

1) ранее известные решения задач по обеспечению ИБ процессов обработки сверхбольших массивов слабоструктурированных данных получены для условий обработки в псевдореальном времени: все такие решения, в основном, ориентированы на методологию «отображения и свертки» (MapReduce) — имеются лишь единичные, не связанные в единый цикл работы, посвященные анализу задач обеспечения ИБ в режиме реального времени или обсуждению различий между ними;

2) необходимо создать комплекс моделей объекта защиты, учитывающих, в отличие от ранее известных, функционирование средств и механизмов обеспечения ИБ как элементов сети систем массового обслуживания, представляющих процесс обработки потока в ЦОД;

3) необходимо создать и использовать для проведения исследований формальную модель угроз ИБ и модель нарушителя ИБ, учитывающую не только функциональные, но и вычислительные возможности нарушителя (по аналогии с используемыми при разработке криптографических алгоритмов и процессов моделями), позволяющие, в отличие от ранее использовавшихся качественных или полуформальных моделей, формулировать корректные утверждения о свойствах алгоритмов защиты и получать их доказательства;

4) необходимо получить решения для комплекса задач обеспечения безопасности информации, реализованные в виде алгоритмов обработки входящего потока и внутренних потоков между элементарными операциями ЦОД, а также процессов взаимодействия средств ЦОД между собой, с владельцем и потребителями данных, обеспечивающих, в отличие от ранее известных, выполнение функций защиты в режиме «жесткого» реального времени.

5. Защищенная инфраструктура больших данных

Частным случаем применения перечисленных выше требований является решение задач обеспечения собственной безопасности ИТИ больших масштабов, в которой формируются столь большие и сложно структурированные массивы данных, связанных с событиями ИБ, что их можно отнести к категории больших данных. Решить такую задачу можно лишь в рамках разработки единой защищенной инфраструктуры обеспечения ИБ технологий больших данных, позволяющей собирать, индексировать, нормализовать, анализировать и совместно использовать относящуюся к ИБ корпорации информацию

защищенным образом. Эта инфраструктура должна быть открытой, гибкой и масштабируемой с точно определенными, стандартизованными форматами входных и выходных данных.

Эта инфраструктура должна быть основана на виртуализованных сетях. Использование виртуальных локальных сетей между ЦОД и виртуальными устройствами в качестве внутренней сети для виртуального хоста, реализующего виртуальные коммутаторы, хорошо подходит для передачи больших данных.

Корпорациям нужно отделить трафик своих обычных пользователей от трафика больших данных, характеризующих ИБ их ИТИ. Поскольку межсетевые экраны должны исследовать каждый проходящий через них пакет для каждой сессии путем поддержки перемещения через зашифрованные сетевые туннели только доверенного трафика больших данных и исключения экранов между элементами ЦОД, защищенная инфраструктура может обмениваться данными на требуемых скоростях в режиме реального времени.

Также важно, чтобы в соответствии с признаваемыми корпорацией стандартами были защищены отдельные виртуальные серверы защищенной инфраструктуры обеспечения ИБ технологий больших данных. Хорошей практикой для них является удаление ненужных сервисов (типа FTP), наличие подходящих процессов своевременного управления обновлениями для ПО и ОС, сервисов резервирования и шифрования резервных копий. Нужна и хорошая централизованная система мониторинга самой информации, ее состояния, доступа к ней и событий ИБ, связанных с ней и процессами ее обработки.

В защищенной инфраструктуре должно использоваться и шифрование, особенно когда речь идет об аналитике в больших данных.

Защищенная инфраструктура должна быть интегрирована с существующими средствами и процессами защиты — сначала параллельно с имеющимися соединениями (это касается, например, SIEM-систем), а затем полностью переоснащена под технологии больших данных.

Как и для любой другой инфраструктуры, для разработки, внедрения, анализа и совершенствования защищенной инфраструктуры больших данных, относящихся к ИБ корпорации, должны быть разработаны все процессы, процедуры, документальное обеспечение, кадровые вопросы, включая, например, обучение вопросам обеспечения ИБ персонала ЦОД, и многое другое.

Особо отметим важность использования в рамках защищенной инфраструктуры визуализации больших данных об ИБ обобщенно по всей ИТИ корпорации. В отличие от визуализации физически существующих явлений или данных (например, человеческого тела, географических карт, землетрясения, воронки урагана, цунами и т. п.), визуализация информации об ИБ ИТИ должна представлять некие абстрактные данные о состоянии защищенности ИТИ, что показывает сложность решения данных вопросов. Например, как представить низкоуровневую сетевую информацию или информацию, полученную за длительное время наблюдения от отдельного сетевого устройства для выявления событий ИБ в виде свершившихся атак или только развивающихся на глазах администратора попыток атак на ИТИ организации. Здесь возникают две главные проблемы любой визуализации — сложность, означающая возможность визуализации различных форм представляемых данных, и масштабируемость, означающая возможность визуализации достаточно больших объемов данных с точки зрения как алгоритмической сложности, так и способности отображать информацию больших объемов понятным для человека образом.

Для истинной видимости необходима расширенная аналитика. Подлежащая визуализации информация об ИБ ИТИ организации может быть использована для непосредственного анализа, обнаружения событий ИБ в потоке всех собранных событий (big data), принятия решений в области управления ИБ и ознакомления с этой информацией определенного круга уполномоченных лиц. Поэтому в целях обеспечения ИБ можно выделить много потенциальных областей применения систем визуализации больших данных об ИБ ИТИ, например, решающих следующие важные

задачи защиты корпоративных сетей от атак из внешних сетей типа Интернета или злоупотреблений со стороны инсайдеров:

- распознавание сетевых атак, обнаружение аномальной активности, выявление несанкционированных (мошеннических) операций с информацией и контроль неблагонадежных сотрудников;
- анализ информационных потоков, трассировка перемещения сетевых пакетов по каналам связи, нахождение путей распространения вирусов или функционирования ботнетов;
- контроль доступа ко всем ресурсам ИТИ и обнаружение уязвимостей в ИТИ;
- исследование кода вредоносных программ или вирусов и выявление сигнатур атак в большом объеме информации от пострадавших систем;
- контроль за системными конфигурациями, анализ эффективности настроек СЗИ, изучение взаимодействия работы различных технологий обеспечения ИБ и отдельных СЗИ;
- корреляция обнаруженных событий ИБ и «срез» событий ИБ для выделенного канала связи, сетевого устройства, сетевого протокола, сервиса, приложения и т. п.;
- установление некоторых тенденций (что возможно только на основе больших данных), построение моделей и разработка правил обеспечения ИБ (например, для настройки в межсетевых экранах или системах обнаружения/отражения вторжений);
- выделение приоритетных направлений, требующих немедленного вмешательства для улучшения управления ИБ на уровне всей ИТИ организации;
- визуализация свидетельств компьютерных преступлений для дальнейшего их расследования и т. п.

Заключение

В настоящее время основные сферы применения ИТ больших данных очень разнообразны: фундаментальные и прикладные исследования в области физики элементарных частиц (в частности, международные эксперименты на ускорителях), астрономии (обработка данных о нашей Вселенной, полученных от больших телескопов), биологии (в частности, расшифровка генома человека), экономики (в частности, прогнозирования макроэкономического развития и финансовых рисков); мониторинг окружающей среды, слежение за состоянием атмосферы и прогнозирование погоды; фармацевтическая промышленность (синтез новых химических соединений и создание новых лекарств); многочисленные применения в бизнесе (анализ покупательской активности, спроса, предпочтений потребителей, эффективности рекламы и пр.); предоставление ИТ больших данных в качестве сервисов (готовых функциональных модулей) при реализации других ИТ (в частности, технологий поиска, глубокой аналитической обработки данных с целью выявления скрытых закономерностей); поиск первоисточников информации и извлечение основного содержания (семантики) в сверхбольших массивах документов без непосредственного их прочтения человеком (в частности, в новостных массивах, массивах законодательных актов, массивах научных публикаций (например, для выявления плагиата) и пр.); аналитическая обработка данных о состоянии ИТИ с целью выявления аномалий в функционировании системы, инцидентов ИБ и предотвращения вторжений и т. п.

Но этот широкий спектр применений в настоящее время ограничен из-за нерешенности многих проблем обеспечения ИБ технологии больших данных. Очевидно, что при условии решения хотя бы части наиболее актуальных задач указанная сфера применения способна получить мощный импульс для развития «вширь». В частности, расширение возможно на следующие важные приложения: информационно-аналитическое обеспечение деятельности правительственных, общественных и коммерческих организаций; мониторинг финансовых транзакций, в том числе транзакций по банковским картам, с целью выявления мошеннических операций; высокоэффективное автоматизированное управление технологическими процессами; научная обработка данных клинической медицины, создание индивидуальных лекарственных препаратов, телемедицина и дистанционное предоставление услуг здравоохранения.

Решение проблем обеспечения ИБ позволит, с одной стороны, повысить доверие потребителей к ИТ больших данных, а с другой стороны, существенно снизить риски несанкционированного или нежелательного извлечения информации путем применения технологий интеллектуального анализа данных и аналитической разведки в Интернете, в частности, способно снизить риски сбора данных террористическими группировками, позволит противодействовать легализации доходов, полученных преступным путем, и финансированию терроризма. Для этого в статье поставлена задача обеспечения ИБ для технологий больших данных, намечены основные подходы к ее решению и рассмотрена их специфика, а также кратко описана защищенная инфраструктура для больших данных, относящихся к ИБ ИТИ корпораций. Дальнейшее исследование связано с решением поставленных задач в области обеспечения ИБ для ЦОД, обрабатывающих большие данные.

СПИСОК ЛИТЕРАТУРЫ:

1. *Hornbeck Ryan L.* Batch Versus Streaming: Differentiating Between Tactical and Strategic Big Data Analytics [Электронный ресурс]. URL: http://datatactics.blogspot.ru/2013/02/batch-versus-streaming-differentiating.html (дата обращения: 21.09.2014).
2. *Rajaraman A., Leskovec J., Ullman J. D.* Mining of Massive Datasets. Cambridge University Press, 2011. — 326 p.
3. *Glick B.* Information security is a big data issue [Электронный ресурс]. URL: http://www.computerweekly.com/feature/Information-security-is-a-big-data-issue (дата обращения: 21.09.2014).
4. *Wood P.* How to tackle big data from a security point of view [Электронный ресурс]. URL: http://www.computerweekly.com/feature/How-to-tackle-big-data-from-a-security-point-of-view (дата обращения: 21.09.2014).
5. *Gentry C.* A Fully Homomorphic Encryption Scheme. PhD Dissertation. Stanford University, 2009.
6. *Jin X., Krishnan R., Sandhu R.* A Unified Attribute-Based Access Control Model Covering DAC, MAC and RBAC // Data and Applications Security and Privacy XXVI. Lecture Notes in Computer Science. 2012. T. 7371. P. 41–55.
7. *Phillips J. M.* Locality Sensitive Hashing. University of Utah, 2013.
8. *Simmhan Y. L., Plale B., Gannon D.* A Survey of Data Provenance Techniques. Technical Report IUB-CS-TR618. Computer Science Department, Indiana University. Bloomington.
9. *Parno B., Howell J., Gentry C., Raykova M.* Pinocchio: Nearly Practical Verifiable Computation. In Proceedings of IACR Cryptology ePrint Archive. 2013. P. 279.
10. *Guarino A.* Digital Forensics as a Big Data Challenge // ISSE 2013 Securing Electronic Business Processes. P. 197–203.
11. *Backes M., Fiore D., Reischu R. M.* Verifiable Delegation of Computation on Outsourced Data // Proceedings of the 2013 ACM SIGSAC conference on computer & communications security. P. 863–874.
12. Massively Parallel Processing (DW) — a Technical Reference Guide for Designing Mission-Critical DW Solutions [Электронный ресурс]. URL: http://technet.microsoft.com/en-us/library/hh393582.aspx (дата обращения: 21.09.2014).

REFERENCES:

1. *Hornbeck Ryan L.* Batch Versus Streaming: Differentiating Between Tactical and Strategic Big Data Analytics [Электронный ресурс]. URL: http://datatactics.blogspot.ru/2013/02/batch-versus-streaming-differentiating.html (05.05.2014).
2. *Rajaraman A., Leskovec J., Ullman J. D.* Mining of Massive Datasets. Cambridge University Press, 2011. 326 p.
3. *Glick B.* Information security is a big data issue. URL: http://www.computerweekly.com/feature/Information-security-is-a-big-data-issue (05.05.2014).
4. *Wood P.* How to tackle big data from a security point of view. URL: http://www.computerweekly.com/feature/How-to-tackle-big-data-from-a-security-point-of-view (29.03.2014).
5. *Gentry C.* A Fully Homomorphic Encryption Scheme. PhD Dissertation. September 2009.
6. *Jin X., Krishnan R., Sandhu R.* A Unifed Attribute-Based Access Control Model Covering DAC, MAC and RBAC. Data and Applications Security and Privacy XXVI. Lecture Notes in Computer Science, 2012. Vol. 7371. P. 41–55.
7. *Phillips J. M.* Locality Sensitive Hashing. University of Utah, 2013.
8. *Simmhan Y. L., Plale B., Gannon D.* A Survey of Data Provenance Techniques. Technical Report IUB-CS-TR618. Computer Science Department, Indiana University, Bloomington IN 47405.
9. *Parno B., Howell J., Gentry C., Raykova M.* Pinocchio: Nearly Practical Verifiable Computation. IACR Cryptology ePrint Archive 2013: 279 (2013).
10. *Guarino A.* Digital Forensics as a Big Data Challenge. ISSE 2013 Securing Electronic Business Processes. P. 197–203.
12. *Backes M., Fiore D., Reischu R. M.* Verifiable Delegation of Computation on Outsourced Data. Proceedings of the 2013 ACM SIGSAC conference on computer & communications security. P. 863–874.
13. Massively Parallel Processing (DW) — a Technical Reference Guide for Designing Mission-Critical DW Solutions. URL: http://technet.microsoft.com/en-us/library/hh393582.aspx (05.05.2014).