УДК 681.3

Предотвращение утечек конфиденциальной информации с помощью технологий анализа DLP-систем

Цыбульский Анатолий Сергеевич
Российский экономический университет имени Г. В. Плеханова
студент

Родионова Дарья Фёдоровна
Российский экономический университет имени Г. В. Плеханова
студент

Аннотация
В статье описывается DLP-система в целом, определяются различия между понятиями контекст и контент информации, различия между DLP-системами и программами с функциями DLP. Рассматриваются различные технологии анализа, их достоинства и недостатки. В статье обозначены основные способы применения каждой из технологий, указаны типы документов, которые могут быть распознаны каждой из указанных технологий.
Ключевые слова: DLP, технология анализа, утечка конфиденциальной информации.

The prevention of sensitive information leaks using DLP-systems` analysis technologies

Tsybulskiy Anatoliy Sergeevich
Plekhanov Russian University of Economics
student

Rodionova Daria Fedorovna
Plekhanov Russian University of Economics
student

Abstract
The article contains overall description of the DLP-system; differences between DLP-systems and programs with DLP functions are stated. Several content and context analysis technologies are reviewed and their advantages and disadvantages are defined. Data types that could be detected using these technologies are listed in the article.
Keywords: DLP, analysis technology, leak of sensitive information.

Каждая информационная система подвергается риску утечек конфиденциальной информации, например, коммерческой тайны или личных данных сотрудников и клиентов. Неудивительно, что технологии

предотвращения утечек информации становятся очень популярны. В то же время такой класс технологий обеспечения информационной безопасности являются достаточно сложными для понимания. Как следствие, довольно сложно понять чрезвычайную важность таких инструментов из-за почти десятка разных технологических подходов.

В западной литературе технологии предотвращения утечек информации называют Data Leak Prevention Technologies, наиболее распространена аббревиатура DLP. В этой статье будут рассмотрены принципы работы основных технологий анализа инструментов защиты от утечек данных.

Предшественниками комплексных DLP-технологий были попытки отделов безопасности в крупных компаниях контролировать аппаратные порты на рабочих станциях их сотрудников. Нередко усилия были направлены и на повсеместное шифрование файлов на компьютерах и сетевых дисках. Сейчас компании, занимающиеся исследованиями в сфере информационной безопасности, определяют полноценные DLP комплексы как продукты, которые будучи основанными на некоторых политиках безопасности, определяют, отслеживают и защищают данные при хранении, использовании и копировании, используя технологию глубокого контентного анализа.

Таким образом, можно выделить обязательные ключевые характеристики полноценной DLP-системы:

1. Технология глубокого контентного анализа
2. Управление политиками безопасности
3. Широкий охват контента на различных платформах и локациях

Системы DLP призваны в первую очередь защищать уязвимые данные, например, коммерческую тайну компании или личные данные пользователей. Другой крайне полезной функцией является проводимый в режиме реального времени анализ контента, обращающегося внутри компании.

Принцип действия любой DLP-системы можно кратко и поверхностно описать в четыре композиционных процесса:

1. Перехват передаваемых данных по сети и на рабочих станциях сотрудников.
2. Анализ контекста и контента данных.
3. Применения политик безопасности.
4. Сохранение события в базу данных.

Конечно, некоторые отдельные продукты могут предоставить частичный функционал комплексных DLP систем, но они всегда будут оставаться ограниченными как глубиной и производительностью своего контент-анализа, так и охватываемых в своем диапазоне данных (только по сети или только на рабочей станции).

Рынок DLP поделён между отдельными программами, предоставляющими некоторые базовые функции защиты от утечек ценных (например, контроль за электронной почтой), и полноценными системами предотвращения утечек.

Полноценные DLP-системы включают в себя централизованные консоли управления, механизм создания политик безопасности и контроль за документооборотом, осуществляемый в целях мониторинга и защиты ценных данных. Функционал системы и интерфейс пользователя разработаны для решения проблем бизнес-направленности, предоставляя отделу безопасности информацию и контроль над контентом, с которыми ведётся работа на компьютерах сотрудников компании.

Отдельные программы в свою очередь предоставляют ограниченный функционал детектирования и запрета действий с некоторыми видами информации, но никак не защищают контент и данные, как это делают DLP системы с использованием политик. В некоторых случаях этого может быть достаточно для компаний, не входящих в зону риска утечки ценных данных из-за отсутствия таковых. Управлять такими программами в большинстве случаев способны только специалисты в области информационной безопасности.

Эти отличия важны, так как полноценные DLP-системы решают конкретную бизнес- проблему – предупреждают и предотвращают утечки ценных корпоративных данных. Управлять такими системами могут даже пользователи, далекие от технических специальностей. Нередко за защиту информации могут назначить ответственным служащего судебного ведомства или органов по охране общественного порядка, обученного работе с DLP-комплексом. Для некоторых компаний будет достаточно только лишь грамотно настроенных и изредка актуализируемых политик безопасности внутри DLP-системы.

Из всего вышесказанного становится понятно, что DLP-системы направлены на решение конкретной бизнес-задачи - защита ценных корпоративных данных, а не на ряд других задач информационной безопасности, таких как защита персонального компьютера или сети, решаемых антивирусами. Это также означает, что комплексные DLP системы подходят не только для крупных компаний, но и для среднего-малого бизнеса, потому что большинство современных DLP-решений имеют модульную структуру – можно приобрести только необходимые модули для установки.

Еще одно крайне важное преимущество DLP систем – они крайне эффективны при детектировании ошибочных действий персонала и ошибочных бизнес-процессов (обмен нешифрованными ценными файлами, ошибка в адресате письма и прочее).

Ключевую функция DLP-систем принято называть «понимание контента».Так называют способность системы проводить глубокий анализ контента, используя для этого различные технологии. Однако следует отличать «понимание контента» от «понимания контекста» информации. Соответственно, для того, чтобы разобраться в сути работы DLP-систем необходимо различать такие понятия как контент и контекст.

Обычно для понимания разницы достаточно запомнить ассоциацию: проще всего представить контент как письмо, и контекст - как конверт и прочее окружение вокруг письма.

Контекст письма обычно включает в себя источники, направления, размер, получателей, отправителя, дату, заголовок, формат и прочие метаданные письма. Контент в свою очередь – это содержимое конверта, письмо, текст, рисунки, любые данные в нем.

Технология понимания контента по сути изучает и анализирует содержимое конверта - само письмо. Основным преимуществом этой технологии является возможность избавиться от ограничений по контексту (окружению) при мониторинге данных и действий с ними в режиме онлайн.

Действительно, если нам нужно защитить какой-то массив ценных данных, то мы хотели бы защитить его повсюду, а не просто в каком-то конкретном контексте (в конкретном файле, папке, мессенджере, окружении). То есть существует необходимость защитить ценные данные в «письме», а не его «конверт». Очевидно, такой подход к защите данных труден в исполнении и требует больше времени, чем базовый контекстуальный анализ. Поэтому технология «понимания контента» и является ядром большинства комплексных DLP систем. Для упрощения понимания далее в статье под термином анализ данных понимается термин анализ контента информации.

Безусловно, контекст сам по себе тоже очень полезен и большинство законченных DLP-решений включает в себя не только контентный анализ, но и контекстный - правда уже в качестве добавочной функции. Чаще всего это реализовано в виде анализа бизнес контекста, что позволяет детектировать особо ценные бизнес данные в их режиме хранения, использования или копирования.

Первым шагом при анализе данных является перехват пакета данных и его экстракция. После этого программе необходимо провести структурный анализ контекста перехваченных данных и разобраться в нем. Этот этап легко реализуется в случае обычного email-письма, но в случае анализа двоичных файлов все не так просто – DLP системе необходимо использовать технологии «взлома» файлов (filecracking). Эта технология используется для чтения и анализа файла даже в том случае, если непосредственное содержимое самого файла скрыто. Например, технология может проанализировать информацию из вставленного в MSWord-файл Excel-листа причём Word-файл может быть расположен даже в заархивированной папке. Система способна распаковать архив, прочитать и проанализировать файл Word, выделить данные из таблицы Excel, прочитать их и проанализировать. Другие ситуации могут быть сложнее, например, pdf находящийся в файле формата CAD. Но и такие случаи можно обработать. Большинство представленных на рынке программных продуктов поддерживают до 300 различных типов файлов, вставленный многоуровневый контент других форматов, различные языки, а также средства для обнаружения текста в документах с неопределенным расширением.

Некоторые из продуктов используют инструменты для «взлома» файлов распространяемые бесплатно и имеющие открытый код. Однако самые популярные DLP решения используют собственные или проприетарные возможности для «взлома» файлов.

Некоторые инструменты поддерживают анализ зашифрованных данных, если шифрование было выполнено с ключами восстановления. Большинство инструментов могут идентифицировать стандартное шифрование, используемое в компании, и используют эту возможность как контекстуальное правило для блокировки или пропуска пакета данных, в зависимости от применённых политик.

После получения доступа к данным, DLP-система использует несколько основных средств анализа для детектирования нарушений правил безопасности. Каждое средство анализа имеет свои сильные и слабые стороны.

1. Технология, основанная на правилах/регулярных выражениях:

Самой распространённой техникой анализа контента в DLP продуктах и других инструментах, использующих DLP-возможности, является техника, основанная на правилах или регулярных выражениях. С ее помощью анализируются данные подходящие под определенные правила безопасности. Например, набор из 16 цифр, который подходит под требования проверочной суммы кредитной карты, характерный номер лицензии Windows, особая шифровка документов, представляющих государственную тайну и прочие эталонные документы, которые можно описать регулярными выражениями. Большинство DLP решений вместе с базовыми регулярными выражения используют дополнительные правила анализа. (Например, имя, указанное вместе с адресом, расположенное рядом с номером кредитной карты, несколько вариантов разделителей в выражениях и другие).

Эта техника наиболее эффективна на первом этапе фильтрации, когда необходимо найти элементы данных, которые легко идентифицируются (номера кредитных карт, номера социальных карт, а также записи медицинских страховок и т.д.) Такие правила быстро выполняются и могут быть легко сконфигурированы, тем более, что большинство продуктов поставляются с начальным набором правил и шаблонов. Технология понятна и проста во внедрении в различные программные продукты. Однако эта техника склонна к высокой степени ложноположительных результатов (когда вердикты программы и специалиста безопасности не совпадают), следственно обеспечивает невысокую степень защиты для неструктурированных данных.

2. Лингвистический анализ:

Лингвистический анализ проводится с целью определения тематики текста и отнесения его к какой-либо категории (например, категории «финансы», «кадры» или «государственная тайна»), что позволит в дальнейшем применить определенные политики безопасности к файлу или сообщению, содержащему текст. Соотнесение текста к категории происходит на основе сравнения с многочисленными словарями терминов. В DLP-

системах производится автоматическое определение языка текста, учитывается морфология, транслитерация и опечатки в словах.

3. Точное совпадение данных выгружаемых таблиц:

Техника точного совпадения данных из таблиц собирает цифровые отпечатки с загруженных в DLP-систему таблиц и в дальнейшем детектирует выгрузки из БД, содержащие точные совпадения данных из защищаемых таблиц, по любым каналам.

Например, можно создать политику для детектирования только номеров кредитных карт в клиентской базе, при этом игнорируя данные сотрудников компании, делающих покупки онлайн. Усовершенствованные инструменты точного совпадения данных детектируют комбинации данных, такие как комбинация имени, инициалов и фамилии с номером кредитной или социальной карты. У этой технологии очень низкий уровень ложноположительных результатов. Позволяет к примеру, защитить персональные данные клиентов, игнорируя при этом другие похожие данные, используемые сотрудниками.

4. Точное совпадение файлов:

Эта технология анализирует хеш-сумму и сигнатуры (цифровой отпечаток) файла и ищет другие файлы с полным совпадением хеш-сумм и сигнатур. Эту технологию можно отнести к технологиям контекстуального анализа потому что само содержание файлов не подвергается анализу.Подходит для медиа-файлов и других файлов, для которых нет необходимости анализировать содержание. Сильным преимуществом является то, что технология работает с любыми типами файлов, дает низкий уровень ложноположительных результатов. Помимо точного совпадения в некоторых DLP-системах реализована технология частичного совпадения. Эта технология выбирает либо частичные, либо полные совпадения с охраняемым документом. В связи с этим администратор может настроить политику для защиты персональных документов, и DLP решение будет выбирать либо полное, либо частичное совпадение с загруженным эталонным документом, вплоть до отдельных незначительных совпадений из нескольких предложений. Например, можно загрузить в систему бизнес-план для нового продукта, DLP решение выдаст уведомление в случае, если сотрудник попробует переслать какой-то абзац из этого документа через средство мгновенного обмена сообщениями или мессенджер на рабочей станции или другой канал. Большинство решений основаны на так называемой технологии циклического хэширования, которая работает следующим образом: берется часть общей хэш-суммы файла, смещается заранее определенное количество символов, далее берется следующая часть данных. Этот процесс повторяется до тех пор, пока документ полностью не будет загружен в виде серии накладывающихся друг на друга хэш-сумм. Исходящие данные проходят через эту хэш-технологию, и хэш-суммы сравниваются между собой. Многие продукты используют циклическое хэширование в качестве исходной базы, применяя в добавок к ней более

совершенные технологии лингвистического анализа. Можно использовать для зашиты исходного кода ПО.

5. Графический анализ:

Графический анализ (Optical Character Recognition), позволяет выявить и распознать среди потока данных определенные графические элементы, как например сканы паспортов граждан РФ, бланков ЕГЭ или даже фотографии банковских карт. Распознавание обычно осуществляется с использованием технологии преобразования элементов графического объекта в текст, дополненной элементами машинного обучения на примерах (обучающая выборка, на основе система восстанавливает зависимости и строит алгоритм распознавания). Такой подход позволяет резко снизить количество ложных срабатываний.

В результате написания статьи была изучена литература по теме исследования, проведен анализ предметной области и выделены следующие выводы: технологии анализа, рассмотренные в статье, составляют основу всех современных DLP решений на российском и международном рынке, хотя есть разница в их реализации и комбинации. На рынке DLP-систем присутствуют как иностранные компании, так и российские. Эти системы начали развиваться в связи с ростом потребностей компаний в этой сфере. Яркими представителями, стоящими у истоков развития таких систем на российском рынке, можно назвать компании «Инфосистемы Джет» и InfoWatch. Каждая компания предоставляет свой спектр услуг, функций и возможностей. Комбинируя существующие технологии анализа, а также другие компоненты и возможности, каждая компания может предоставить подходящее DLP-решение.

Библиографический список

1. Ковалёв А. Как выбрать DLP-систему? // PCWeek/RE «Компьютерная неделя»). № 42 (744). 9 -15 ноября 2010 г.
2. Ерыгин А.В. Анализ эффективности систем предотвращения утечек конфиденциальной информации из локальных сетей // Вестник СибАДИ. 2011. №2 (20). С.47-52.
3. Барабанов А.В., Гришин М.И., Марков А.С., Цирлов В.Л. Формирование требований по безопасности информации к DLP-системам // Вопросы радиоэлектроники. 2013. №2. С. 67-76.