ПОДХОД К АНАЛИЗУ И ОЦЕНКЕ ЗАЩИЩЕННОСТИ СИСТЕМ УПРАВЛЕНИЯ БОЛЬШИМИ ДАННЫМИ

Полтавцева М. А.[1], Зегжда Д. П.[2]

DOI: 10.21681/2311-3456-2025-5-109-118

Цель исследования: *разработка подхода к анализу и оценке защищенности систем управления большими данными, с учетом технологических особенностей данного класса решений, отличающих их от традиционных облачных систем обработки данных.*

Метод(ы) исследования: *в работе проводится анализ особенностей целевых систем, а также предлагаемых исследователями методов сбора данных и оценки защищенности. Выделяются их недостатки в контексте современных требований. Предлагается использовать подход к сбору данных и моделированию целевой системы с использованием теоретико-множественной агрегатной модели данных, а также интеграция модифицированной оценки NIST контроля доступа в системах больших данных и авторской оценки на основе учета грануляции данных и доверия к узлам-обработчикам.*

Результат(ы) исследования: *в результате работы были сформулированы технологические особенности целевых систем с точки зрения оценки защищенности, такие как распределенность, гетерогенность (мультимодельность), сложный жизненный цикл данных. Анализ научных работ показал, с одной стороны, интерес исследователей к задаче оценки защищенности систем управления большими данными, а с другой – отсутствие оценок, предложенных для целевого класса систем. Авторами сформированы требования к оценке защищенности к системам управления большими данными как к специфическому компоненту современных информационных систем. Также предложен новый метод оценки защищенности, впервые учитывающий специфические свойства систем управления большими данными. Предлагаемый метод, дополнительно к ранее предложенным оценкам, учитывает недостатки контроля доступа вызванные различной грануляцией данных в компонентах целевой системы, а также большое число доверенных пользователей, и, как следствие, необходимость обработки конфиденциальных данных либо на доверенных узлах, либо в скрытом (обфусцированном или зашифрованном) виде. Предложенная оценка является нормированной, может быть детализирована до оценки каждого конкретного инструмента обработки данных, легко расширена или встроена в более высокоуровневые оценки. Показана достоверность и возможность практического применения предложенной оценки путем разработки программного прототипа на основе ранее известных и апробированных программных решений.*

Научная новизна: *новизна заключается в авторском методе оценки защищенности систем управления большими данными, отличающимся впервые предложенным учетом недостатков контроля доступа вызванных различной грануляцией данных и учетом доверия к отдельным узлам обработчикам данных.*

Ключевые слова: *большие данные, информационная безопасность, оценка защищенности, грануляция данных, контроль доступа, доверие.*

Введение

Сегодня в результате развития цифровых технологий практически во всех отраслях широко используется сбор, обработка, хранение и использование больших данных. Технологически, формируется новый класс информационных систем, в составе которых вместо классического сервера систем управления базами данных (СУБД), для хранения и обработки информации используется сочетание разнородных компонентов: набор СУБД различного типа, инструменты потоковой обработки информации, иное программное обеспечение различного типа.

Такие системы сталкиваются с большим количеством угроз и проблем безопасности, связанных с особенностями их построения [1]. В свою очередь недостатками систем защиты таких решений, из-за их технологических особенностей, являются децентрализованные и несогласованные механизмы контроля доступа отдельных инструментов хранения и обработки данных, высокое доверие между инструментами, сложность согласованного администрирования.

Фундаментальными причинами проблемы безопасности в рассматриваемом классе систем является, в первую очередь различная грануляция данных в инструментах их обработки [2]. Эта особенность, в свою очередь, порождает различия в используемых моделях контроля доступа и их реализациях в разных компонентах системы, что, в свою очередь, зачастую является причиной несанкционированного доступа к данным.

1 Полтавцева Мария Анатольевна, доктор технических наук, доцент, Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого». Санкт-Петербург, Россия. E-mail: poltavtseva@ibks.spbstu.ru

2 Зегжда Дмитрий Петрович, член корреспондент РАН, доктор технических наук., профессор, Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого». Санкт-Петербург, Россия. E-mail: zegzhda_dp@spbstu.ru

Целью данной работы является создание подхода к анализу и последующей оценке защищенности систем управления большими данными с учетом ключевых особенностей, отличающих данный класс решений от других типов информационных систем.

Степень разработанности проблемы

Проблемой безопасности в системах обработки и хранения больших данных занимаются многие исследователи. Нужно отметить, что на данный момент большая часть работ посвящена контролю доступа в таких системах. Также исследователи отмечают задачи сбора данных и оценки защищенности, однако эти области являются в значительной степени менее исследованными [3].

Наибольшая часть современных научных работ посвящена мониторингу и анализу управления большими данными в облачных системах. В частности, можно отметить исследования посвященные безопасности облачных систем и процессов разработки [4]. Однако распределенные гетерогенные системы больших данных исследователи в этой области не рассматривают, ограничиваясь рамками одного центра обработки данных и одной экосистемы инструментов обработки. Ряд работ посвящены использованию blockchain и технологиям распределенных реестров. Это работы как по управлению большими данными при помощи данной технологии [5], так и по мониторингу отдельных технологических сред, в частности – интернета вещей [6], промышленных решений [7]. Недостатком этих работ также является их ориентированность на гомогенную среду, в крайнем случае – экосистему и решения одного производителя. Предложенное в [8] решение пока не имеет известных аналогов для гетерогенных систем.

С точки зрения оценки защищенности сегодня работ также крайне мало, в том числе потому, что такая оценка должна базироваться на данных, полученных в результате анализа системы. А задача сбора данных для такого анализа, типового решения, как показано выше, еще не имеет. Как правило, исследователи оценивают безопасность на уровне информационных систем в целом. Для больших данных предлагается оценка безопасности на основе теории принятия решений [9], различные подходы по оценке рисков (например, [10]). Проблема безопасности и оценки защищенности гетерогенных архитектур больших данных является фокусом исследования в работе [11], однако конкретного решения ее авторы не предлагают. Необходимость оценки защищенности через анализ контроля доступа авторы рассматривают и обосновывают в работе [12], однако сама оценка не приводится. А оценка из работы [13] также не учитывает иных свойств данного класса систем, кроме необходимости согласования контроля доступа.

С точки зрения оценки защищенности гетерогенные системы обработки и хранения больших данных являются частью информационных систем, и оценки защищенности информационных систем релевантны в их отношении [14]. В свою очередь, исследования в области оценки защищенности информационных систем можно отнести к следующим категориям. Во-первых, это работы, которые рассматривают отдельные аспекты защищенности систем и технологий (например, [15]). Во-вторых, работы, посвященные конкретным инфраструктурам [16], включая оценку рисков безопасности и облачные инфраструктуры [17]. Здесь стоит отдельно отметить работу [18]. В ней авторы анализируют гетерогенные архитектуры, однако специфические проблемы гетерогенных систем обработки и хранения данных ими также не рассматриваются.

Таким образом, задача оценки защищенности гетерогенных систем управления большими данными, включая как сбор и анализ информации, так и вычисление самой метрики защищенности, является высоко актуальной.

Особенности гетерогенных систем управления большими данными как объекта защиты

Гетерогенные системы обработки и хранения больших данных обладают рядом ключевых особенностей [19]. Эти особенности являются их нативными технологическими чертами систем управления большими данными и отличают их как от классических систем управления базами данных, так и от более общих классов решений (например, информационных систем). Ключевыми технологическими чертами нового класса, гетерогенных систем управления большими данными, являются:

1. Распределенная в широком смысле среда обработки.
2. Использование нескольких разнородных (основанных на разных моделях данных) инструментов обработки данных.
3. Единый жизненный цикл разнородных данных.

В рамках рассматриваемых систем данные обрабатываются, как правило, в рамках нескольких связанных СУБД функционирующих на основе различных моделей данных. При этом, в силу разнородности и сложности задач, каждая такая система является самостоятельным компонентом со своими пользователями и жизненным циклом данных [1]. Как правило, все или часть таких СУБД развернуты отдельно друг от друга на географически распределенных серверах или центрах обработки данных. В результате в организации используется распределенная в широком смысле среда обработки данных, вся информация в которой связана в единый жизненный цикл. Результаты обработки или сырые данные

Рис. 1. Гетерогенные системы обработки и хранения больших данных

с одного этапа или из одной системы являются входными данными для другой и служат целевой информацией для ее пользователей. В результате одни и те же данные обрабатываются на различных узлах под управлением различных СУБД в распределенной корпоративной среде (рисунок 1).

С точки зрения решения задач безопасности это приводит к необходимости учета следующих моментов при оценке защищенности:

1. Каждый отдельный инструмент хранения и обработки данных нуждается в собственной оценке и анализе защищенности [20].
2. Политики безопасности в рамках отдельных инструментов должны быть согласованы, отсутствие такого согласования или его несоответствия, вызванные технологической невозможностью должной грануляции данных, должны учитываться как потенциальные уязвимости [21].
3. Доверие в отношении узлов – обработчиков данных должно оцениваться и обработка открытых данных на узлах с низким уровнем доверия (большим числом доверенных привилегированных пользователей или иными факторами риска) должна быть минимизирована [22].
4. Необходима возможность получения интегрированной оценки защищенности для информационной системы в целом.

В этих условиях основной задачей становится сбор данных и оценка защищенности, которая, с одной стороны, учитывает безопасность отдельного инструмента, с другой – технологические особенности целевого класса систем (систем управления большими данными), и с третьей - может быть интегрирована в оценку защищенности информационной системы организации в целом. В оценке защищенности таких систем требуется учесть несколько факторов.

Во-первых, это недостатки системы контроля доступа. Как показано в работах [2, 12], при использовании инструментов с различной грануляцией данных неизбежно возникают проблемы согласования доступа. Также характерная особенность таких систем — это повышение привилегий отдельных пользователей из-за недостаточной грануляции слабо структурированных данных для обеспечения бизнес-процессов.

Во-вторых, это динамичность компонентов обработки и хранения данных. Указанные организации как правило имеют большой объем legacy, в том числе в области организации бизнес-процессов, что вызывает дублирование данных. Организационные изменения и работа с персоналом также приводят к возникновению новых этапов и шагов в обработке информации. Отслеживание и учет таких этапов это также важная задача.

В любом случае, оценка защищенности (в том числе для систем обработки и хранения больших данных) базируется на результатах сбора и анализа данных в процессе мониторинга и определении на их основе численных метрик.

Оценка защищенности гетерогенных систем больших данных

Сбор и анализ данных для оценки защищенности

Сбор и анализ информации в гетерогенных системах обработки и хранения больших данных для решения задачи оценки защищенности должен обеспечивать информированность о движении потоков данных (по крайней мере, между инструментами обработки) и сведениями о состоянии и согласованности контроля доступа в инструментах обработки и хранения информации.

Для решения этой задачи была использована система мониторинга и моделирования процессов обработки данных в системах управления большими данными на основе распределенного реестра [12]. Моделирование процесса обработки данных, на основе реальных потоков данных в целевой системе, позволяет установить изменения в процессах обработки и хранения данных для оценки сведений о движении фрагментов данных. Над полученными данными авторским коллективом ранее был разработан метод анализа политик безопасности, с целью поиска ошибок и уязвимостей. Анализ политики безопасности решает несколько задач:

1. Проверка соответствия частных политик безопасности на инструментах обработки и хранения данных политике верхнего уровня.
2. Обнаружение возможных каналов логического вывода со стороны получателей данных.
3. Итерационный поиск оптимальной политики безопасности, соответствующей принципу минимального доверия, с учетом технологических ограничений.

В результате может быть установлено соответствие реализации политики безопасности и ее оптимальной конфигурации, а также выявлены неустранимые недостатки, которые должны быть компенсированы иными мерами. Общая схема сбора данных и их применения на последующих шагах по расчету оценки защищенности приведена на рисунке 2.

Рис. 2. Схема сбора данных и оценки защищенности в гетерогенных системах обработки и хранения больших данных

На первых этапах выполняется сбор данных на узлах-обработчиках, на уровне отдельных инструментов, и сохранение в цепочке распределенного реестра. Собираемые данные включают информацию об узлах системы управления большими данными, типах фрагментов данных и выполняемых операциях над ними. Технологически сбор данных может быть основан на традиционных технологиях, однако использование распределенного реестра позволяет гарантировать целостность информации об узлах и выполняемых операциях, поступающих для дальнейшей оценки защищенности.

Затем осуществляется анализ полученных цепочек распределенного реестра уже в блоке управления безопасностью, включая извлечение данных о последовательностях операций над каждым фрагментом больших данных. На основе извлеченных данных блоком моделирования выполняется построение модели системы сбора, хранения и обработки больших данных в виде графа обработки информации в целевой гетерогенной системе. Вершинами такого графа являются операции над данными, а ребрами – передаваемые фрагменты данных.

Для согласования логических представлениий инструментов обработки данных (моделей данных) при составлении модели системы обработки информации в целом, на основе аппарата теории множеств была предложена агрегатная модель данных [23].

Пусть $d_i \in D$ – фрагмент данных, передаваемый между инструментами и/или обработки информации. Тогда каждый i-й фрагмент, точнее – тип фрагмента, представляется кортежем ключ – значение $d_i = <Key, Val>$, где ключом является уникальный идентификатор типа фрагмента данных. Значение фрагмента данных – содержащаяся в нем информация, включая, в некоторых случаях, пустое множество: $d_i.Value = \{Val, \{d_{ij} | d_i \in \{\emptyset, d\}\}$. Каждый элемент данных может быть атомарным, и тогда ключ является произвольным UID внутри системы. Либо же элемент данных может представлять собой агрегацию других элементов (записи в кортеже, кортежи в таблице, документы в коллекции и т.д.). Идентификатор (ключ) части составного фрагмента данных представляет собой конкатенацию собственного UID и UID обещающего фрагмента: $d_i.Key = \overline{Key} = (k_n, \ldots, k_1)$.

В рамках модели выделяются основные операции над фрагментами данных, такие как:

Создание фрагмента:

$$Create: (\{d\}, Key) \longrightarrow d_i = \{Key, Value\}.$$

Удаление фрагмента: $Delete: (d_i) \longrightarrow \emptyset$.

Включение одного фрагмента в другой:

$$Incl{:}(d_i,\{d\}) \longrightarrow d_i`.$$

Исключение части составного фрагмента:

$$Exception\ Extr{:}(d_i`,\{Key_j\}) \longrightarrow (d_i,d_j).$$

Изменение фрагмента данных, в общем случае структурное (например, сортировка) или семантическое (например, вычисление над ним некоторой функции или иное преобразование): $Transform{:}(d_i) \longrightarrow d_j$.

С точки зрения безопасности и оценки защищенности все приведенные операции разделяются на те, которые могут выполняться без доступа к семантике данных на текущей технологической платформе и те, которые требуют такого доступа.

Без доступа к семантике, как правило, выполняются операции удаления, включения и, часто, исключения фрагментов данных. Это обусловлено тем, что при удалении чтение содержимого фрагмента не требуется, при включении фрагментов также не нужен доступ к значению данных, а только к синтетическому ключу-идентификатору, а при исключении фрагментов большинство систем также позволяют извлечь часть не оперируя к целому. Здесь безусловно есть некоторые нюансы, связанные, в частности, с механизмами ускорения доступа в системах управления базами данных. Например, для выполнения запроса требуется сравнение критерия запроса со значениями индекса, и таким образом, происходит обращение к семантике. Однако на сегодня известны технологии сквозного шифрования для устранения этой проблемы и для NoSQL решений на базе более простых моделей задача может решаться проще. Поэтому операция исключения в каждом отдельном случае может как требовать доступа к открытой семантике данных, так и быть реализована без нее.

Операции создания и трансформации фрагментов данных очевидным образом требуют доступа к семантике сведений и должны выполняться на доверенных узлах или в защищенном режиме, т.е. над зашифрованными или обфусцированными данными.

Метод оценки защищенности

Определение оценки защищенности (нижний блок на рисунке 2) можно разделить на два этапа, которые могут выполняться параллельно. Первый этап заключается в анализе реализованных политик безопасности в гетерогенной системе сбора и обработки больших данных. На этом этапе для гетерогенных систем обработки и хранения больших данных целесообразно применение двух типов анализа. Во-первых, анализ и оценка защищенности на основе подхода NIST, детально описанный в [13, 24]. Во-вторых, детальный анализ политик безопасности гетерогенных мульти модельных систем, с учетом технологических ограничений и разницы в грануляции данных [12].

Анализ согласно рекомендациям NIST проводится на основе цепочки доверия, в которой выделяются источник данных, ведущая и ведомая системы [24]. Цепочка доверия формируется от источника данных к ведущей системе через соглашение о безопасности, а от ведущей системы к ведомой через список доверенных систем, политику контроля доступа и словарь атрибутов. Обратная связь от ведомой системы к ведущей формируется через локальную политику ведомой системы и словарь атрибутов, тем самым формируя невозможность неавторизованного доступа без участия обеих систем. Ситуация, когда реальная реализация контроля доступа шире, чем должна быть на основе модели, называется нарушением. При нарушении оценка безопасности рассматриваемой системы уменьшается, тем больше, чем выше ее класс секретности. Порядок расчета такой оценки Ecn приведен на рисунке 3, где n и n'– число правд доступа реализованных с нарушениями текущее и итоговое.

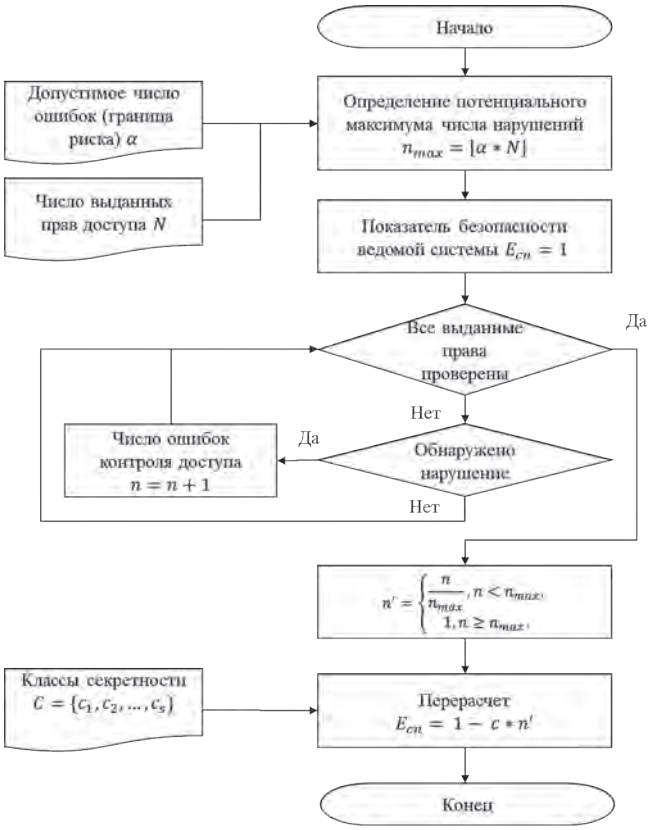

Рис. 3. Алгоритм расчета оценки контроля доступа

При детальном анализе политики безопасности на основе полученной в результате сбора данных модели потоков данных конкретной системы последовательно проводится две оценки. Анализ соответствия политике безопасности верхнего уровня заключается в оценке соответствия общей политике безопасности

политик безопасности реализованных на узлах обработки данных. В результате него формируется оценка $Ec_{,i}$. В наиболее простом случае эта оценка может быть бинарной для узла $Ec_{,i} \in \{0,1\}$, где 0 – не выполняется, 1 – выполняется. Наиболее рациональной представляется следующая оценка:

$$Ec_i = \frac{n_i}{N} \quad (1)$$

где n_i – число отклонений от правил политики безопасности на узле i, а N – общее число правил политики. Полученное значение $Ec_{,i} \in [0,1]$. При этом не учитывается объем отклонений, корректная оценка которого сегодня представляется сложной задачей. Оценка Ec_i является более детальной относительно оценки по NIST (Ec_n).

На втором этапе проводится анализ узлов, выполняющих операции над данными, с учетом степени их доверенности. Цель этой части оценки – выявление операций над открытыми данными на узлах обработки и хранения больших данных, не обладающих должным уровнем доверия. При ее проведении фактически выполняется правило «*все операции с открытыми данными должны выполняться только на доверенных узлах системы*».

Для проведения такого анализа используется отображение графа обработки данных на узлы обработки в модели системы. Оценка самого доверия к узлам при этом может проводится стандартными общеизвестными способами, уровень доверия владелец системы может установить самостоятельно на основ критичности или степени конфиденциальности обрабатываемых данных. Входными данными для дальнейшего расчета является статус узлов: условно «доверенный» или «не доверенный» с точки зрения возможности обработки открытых данных. При анализе узлов системы, выполняющих операции над данными, с учетом степени их доверенности, режим обработки данных считается безопасным в том случае, если обработка данных осуществляется в зашифрованном виде на уровне пользователя либо же данные обфусцированы, то есть, недоступны для злоумышленника, имеющего доступ к узлу – обработчику в том числе, на уровне администратора узла.

Оценка защищенности с учетом доверия в отношении операций по обработке данных рассчитывается как нормированная оценка безопасности операций для каждого i-го узла системы:

$$Eo_i = \frac{\sum_{j=1}^{M} s_j}{M}, \quad (2)$$

где M – число различных типов операций, s_j – признак безопасности j-й операции. Для обработки в безопасном режиме s_i = 1, для обработки в небезопасном режиме s_j = 0. Таким образом, оценка $Eo_i \in [0,1]$ показывает соотношение безопасных и небезопасных операций. На безопасном узле Eo_i = 1. Если Eo_i = 0 – это значит все операции выполняются в небезопасном режиме.

Для системы в целом может быть вычислена суммарная нормированная оценка на основе формул (1) и (2) вида:

$$\mathrm{Es} = \frac{\sum_{i=1}^{N}(Ec_i + Eo_i)}{2N}. \quad (3)$$

Также одним из требований к оценке защищенности была необходимость ее интеграции с другими оценками, характеризующими информационную систему в целом. Так как обе разработанные оценки Ec и Eo являются нормированными, прочие характеристики защищенности при необходимости вводятся дополнительно владельцем системы также как нормированная метрика $E_{ex} \in [0,1]$.

Итоговая оценка защищенности может рассчитывается согласно формуле:

$$E = \frac{E_{ex} + Es}{3}, \quad (4)$$

где E_{ex} – внешняя оценка, E_s – оценка защищенности на основе анализа контроля доступа и безопасности операций, отражающая специфические особенности и уязвимости системы. Полученная оценка является нормализованной, при необходимости детализируется и может входить в более высокоуровневые оценки, включая, при необходимости, показатель состояния технической защиты информации[3]. В формулы (3) и (4) также могут быть добавлены веса для акцентирования различных аспектов функционирования системы управления большими данными.

Стоит отметить, что предложенная оценка защищенности распределенных гетерогенных систем управления большими данными включает не только оценки, полученные известными ранее способами, но и составляющую, определяемую соответствием процессов сбора, хранения и обработки больших данных политике безопасности с учетом различий в структуризации данных на различных узлах целевой системы. Предложенный метод расчета оценки защищенности позволяет как детализировать ее до отдельных узлов, определяя наиболее уязвимые компоненты системы, так и получить общее усредненное значение для сравнения различных конфигураций системы, ее эволюционных версий и иных задач.

Реализация предложенных решений

Для реализации приведенного метода в рамках технологии обеспечения защищенности систем обработки и хранения больших данных была разработана

3 Методика оценки показателя состояния технической защиты информации и обеспечения безопасности значимых объектов критической информационной инфраструктуры Российской Федерации. Методический документ от 2 мая 2024 г. Утвержден ФСТЭК 2 мая 2024 г.

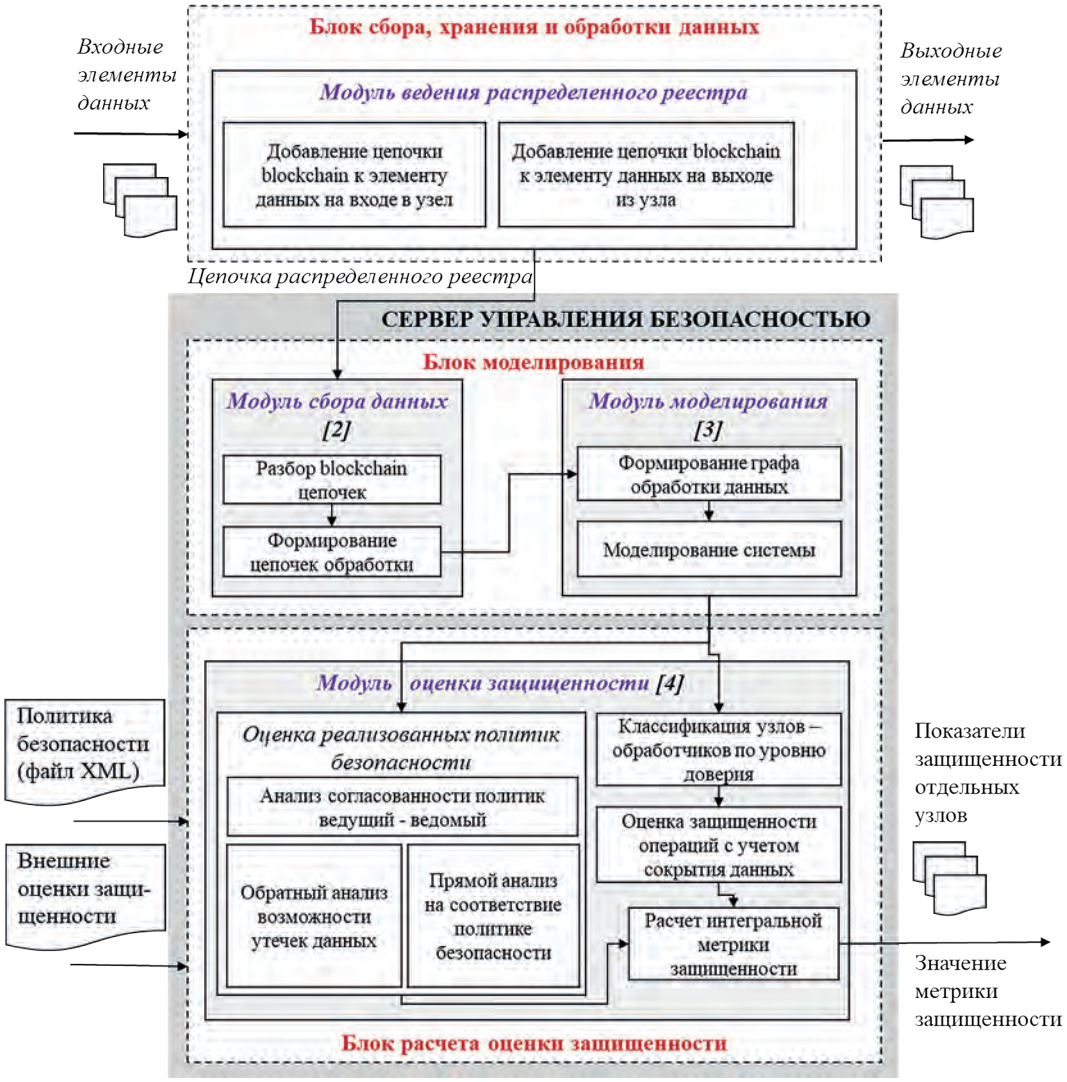

Рис. 4. Система оценки защищенности систем управления большими данными

архитектура системы безопасности и программный прототип для оценки защищенности. Программный прототип реализован на языке python с применением технологии blockchain на этапе сбора и моделирования данных[4].

На узлах – обработчиках данных, на которых функционируют отдельные инструменты гетерогенной обработки информации (PostgreSQL, CassandraDB, MongoDB) и реализуется сложный жизненный цикл обработки данных реализуются программные агенты, собирающие сведения для построения модели системы на основе потоков данных. Центральный сервер управления безопасностью осуществляет сбор и моделирование данных, анализ политики безопасности и расчет оценок (рисунок 4).

Управление указанными фреймворками осуществляется через консоль, на входе принимается json – файлы конфигурации, включая политику безопасности верхнего уровня и внешние оценки безопасности. Влияние на производительность целевой системы со стороны фреймворка сбора данных показывает увеличение нагрузки на систему в пределах 4–10 % [8], вычисление оценки защищенности и работа с политикой безопасности выполняется сервером управления безопасностью не влияет на производительность целевой системы.

4 Свидетельство о государственной регистрации программы для ЭВМ № 2024688201 Российская Федерация. Программа гранулированного аудита в гетерогенных распределенных системах обработки и хранения больших данных: № 2024686659: заявл. 08.11.2024: опубл. 26.11.2024 / М. А. Полтавцева, М. О. Калинин; заявитель Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого». – EDN NEJLCQ
Свидетельство о государственной регистрации программы для ЭВМ № 2024687565 Российская Федерация. Программа автоматического анализа политик безопасности и контроля доступа в системах обработки и хранения больших данных: № 2024686731: заявл. 08.11.2024: опубл. 20.11.2024 / М. А. Полтавцева, М. О. Калинин; заявитель федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого». – EDN CZTMNY

Заключение

При решении поставленной задачи были сформулированы особенности систем управления большими данными, которые отличают их от других программных решений: гетерогенность, географическая распределенность и единый жизненный цикл данных в рамках экосистемы в целом. Эти особенности являются отличительными чертами целевых систем, прямо влияющими на их безопасность и должны быть приняты во внимание при формировании оценки защищенности.

В основе сбора и подготовки данных при оценке защищенности, для отражения особенностей гетерогенных мультимодельных систем управления большими данными, предлагается использовать ранее разработанную систему на основе технологий распределенного реестра. Построенная в результате сбора информации о движении фрагментов данных в целевой системе модель позволяет оценить выполнение политики безопасности не только в формате ведущей - ведомой систем хранения, но и учесть различную грануляцию данных на протяжении их жизненного цикла и связанные с этим возможные нарушения политики безопасности системы.

При расчете оценки защищенности в части, специфической для целевых систем управления большими данными, предложенный новый метод учитывает несколько факторов. Это согласование политик безопасности отдельных инструментов на основе оценки грануляции и подхода NIST и результат анализа узлов, выполняющих операции над данными с точки зрения доверия. В итоге формируется нормированная интегральная оценка, включающая все приведенные аспекты. За счет нормализации предложенная оценка может быть легко интегрирована с иными техниками и оценками защищенности, как применяемыми к целевой системе, так и на более высоком уровне.

Сформированная архитектура программного прототипа и модуль оценки защищенности опираются на использование ранее разработанного фреймворка, апробированы на практике и позволяют рассчитать оценку без дополнительной нагрузки на производительность, за исключением аспекта сбора данных при мониторинге и построении модели целевой системы.

Предложенный метод и система оценки защищенности систем управления большими данными позволяют достичь повышения безопасности целевых систем, в том числе, за счет естественной детализации компонентов оценки до отдельного инструмента обработки данных. Повышение гарантированности выполнения требований безопасности достигается за счет использования распределенного реестра при сборе данных об узлах и операциях над данными в системе сбора, хранения и обработки больших данных, не позволяющего осуществить подмену собранных данных. Автоматизация работы по сбору данных о системе и о процессах обработки данных, а также автоматизация анализа политик безопасности и расчета оценки защищенности позволяют сократить временные и ресурсные затраты на оценку защищенности.

Как дальнейшие направлениями развития данной области можно обозначить повышение качества мониторинга и сбора данных об информационных потоках в системе управления большими данными и повышение степени интеграции предложенных оценок с категориями конфиденциальности данных, принятыми в конкретной целевой организации.

Исследование выполнено за счет гранта Российского научного фонда № 23-11-20003.
https://rscf.ru/project/23-11-20003/, грант Санкт-Петербургского научного фонда
(Соглашение №23-11-20003 о предоставлении регионального гранта).

Литература

1. Минзов А. С., Невский А. Ю., Баронов О. Р. Безопасность персональных данных: новый взгляд на старую проблему // Вопросы кибербезопасности. 2022. №. 4 (50). С. 2–12. DOI:10.21681/2311-3456-2022-4-2-12
2. Colombo P., Ferrari E. Access control technologies for Big Data management systems: literature review and future trends // Cybersecurity. 2019. Т. 2. №. 1. С. 1–13.
3. Rafiq F. et al. Privacy Prevention of Big Data Applications: A Systematic Literature Review // SAGE Open. T.12(2). DOI: 10.1177/21582440221096445
4. Markov A. S., Varenitca V. V., Arustamyan S. S. Topical Issues in the Implementation of Secure Software Development Processes // Proceedings of the International Conference on Information Processes and Systems Development and Quality Assurance IPSQDA-2023 (March 22–24, 2023, St. Petersburg Russia). IEEE. 2023. С. 48–54.
5. Alhazmi H. E., Eassa F. E., Sandokji S. M. Towards Big Data Security Framework by Leveraging Fragmentation and Blockchain Technology // IEEE Access. 2022. Т. 10. С. 10768–10782. DOI: 10.1109/ACCESS.2022.3144632.
6. Wang T. et al. Edge-based auditing method for data security in resource-constrained Internet of Things // Journal of Systems Architecture. 2021. Т. 114. С.1–10. DOI: 10.1016/j.sysarc.2020.101971.

7. Stodt, J. at al. Security Audit of a Blockchain-Based Industrial Application Platform // Algorithms. 2021. Т. 14(4), 121 с. DOI:10.3390/a14040121
8. Kalinin M., Poltavtseva M., Zegzhda D. Ensuring the Big Data Traceability in Heterogeneous Data Systems // 2023 International Russian Automation Conference (RusAutoCon). Sochi, Russian Federation. 2023. С. 775-780. DOI: 10.1109/RusAutoCon58002.2023.10272905.
9. Attaallah A. et al. Analyzing the Big Data Security Through a Unified Decision-Making Approach // Intelligent Automation & Soft Computing, 2022. Т. 32(2). С. 1071–1088. DOI: 10.32604/iasc.2022.022569.
10. Yang M. Information security risk management model for big data // Advances in Multimedia 2022. Т.1 С. 1-10 DOI: 10.1155/2022/3383251.
11. Theodorakopoulos L., Theodoropoulou A., Stamatiou Y. A State-of-the-Art Review in Big Data Management Engineering: Real-Life Case Studies, Challenges, and Future Research Directions // Eng. 2024. Т. 5(3). С. 1266–1297. DOI: 10.3390/eng5030068.
12. Kalinin M., Poltavtseva M. Big Data Security Evaluation by Bidirectional Analysis of Access Control Policy // 2024 International Russian Smart Industry Conference (SmartIndustryCon). Sochi, Russian Federation. 2024. С. 98–103. DOI: 10.1109/SmartIndustryCon61328.2024.10515459.
13. Poltavtseva, M. A., Zaitseva, V. V., Ivanov, D. V. Assessing the Security of Big Data Systems // Aut. Control Comp. Sci. 2024. Т. 58. С. 1352–1364. DOI: 10.3103/S0146411624701025.
14. Dhillon G., Smith K., Dissanayaka I. Information systems security research agenda: Exploring the gap between research and practice // The Journal of Strategic Information Systems. 2021. Т. 30. № 4. DOI: 10.1016/j.jsis.2021.101693.
15. Костогрызов А. И. Методические положения по вероятностном прогнозированию качества функционирования информационных систем ч. 1–3 / А. И. Костогрызов, А. А. Нистратов, П. Е. Голосов // Вопросы кибербезопасности. 2025. № 2(66). С. 2–19. DOI: 10.21681/2311-3456-2025-2-2-19.
16. Оценка уязвимостей автоматизированных систем с применением теории вероятностей, распределения Стьюдента и нормальных случайных величин / И. В. Атласов, А. О. Ефимов, Е. А. Рогозин, А. С. Черкасова // Вопросы кибербезопасности. 2025. № 2(66). С. 124–131. DOI: 10.21681/2311-3456-2025-2-124-131.
17. Ali, T., Al-Khalidi, M., Al-Zaidi, R. Information Security Risk Assessment Methods in Cloud Computing: Comprehensive Review // Journal of Computer Information Systems. 2024. С. 1–28. DOI: 10.1080/08874417.2024.2329985.
18. Крюков Р. О., Федорченко Е. В., Котенко И. В., Новикова Е. С., Зима В. М. Оценивание защищенности гетерогенных инфраструктур на основе графов атак с использованием баз данных NVD и MITRE ATT & CK / Р. О Крюков, Е. В Федорченко, И. В. Котенко, Е. С Новикова, В. М. Зима / Информационно-управляющие системы. 2024. Т.2., С. 39-50. DOI: 10.31799/1684-8853-2024-2-39-50.
19. Wang J. et al. Big data service architecture: a survey // Journal of Internet Technology. 2020. Т. 21. №. 2. С. 393–405.
20. Omotunde H., Ahmed M. A comprehensive review of security measures in database systems: Assessing authentication, access control, and beyond // Mesopotamian Journal of CyberSecurity. 2023. Т. 2023. С. 115–133.
21. El Ahdab L. et al. Unified Models and Framework for Querying Distributed Data Across Polystores //International Conference on Research Challenges in Information Science. Cham: Springer Nature Switzerland. 2024. С. 3–18.
22. Wang S. et al. Data privacy and cybersecurity challenges in the digital transformation of the banking sector // Computers & security. 2024. Т. 147.
23. Poltavtseva M., Aleksandrova E., Izotova O. Data modeling for consistent access control in heterogeneous big data systems // 2024 Ivannikov Memorial Workshop (IVMEM). Velikiy Novgorod, Russian Federation. 2024. С. 42–48. DOI: 10.1109/IVMEM63006.2024.10659707
24. Hu V. C. et al. An access control scheme for big data processing // 10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing. IEEE. 2014. С. 1–7.

AN APPROACH TO ANALYZING AND EVALUATING BIG DATA MANAGEMENT SYSTEMS SECURITY

Poltavtseva M. A.,[5], Zegzhda D. P.[6]

Keywords: *big data, information security, security assessment, data granulation, access control, trustworthiness.*

Purpose of the study: *development of an approach to analyzing and evaluating the security of big data management systems, taking into account the technological features of this class of solutions that distinguish them from traditional cloud data processing systems.*

Methods of research: *the paper analyzes the features of target systems, as well as the methods of data collection and security assessment proposed by another researchers. Their disadvantages are highlighted in the context of modern requirements. It is proposed to use an approach to data collection and modeling of the target system using the aggregate data model based on set theory, as well as the integration of a modified NIST assessment of access control in big data systems and an author's assessment, based on data granulation and trust in processing nodes.*

Result(s): *as the result of the work, the technological features of the target systems were formulated in terms of security assessment. They are distribution, heterogeneity (multimodality), and a complex data lifecycle. The analysis of scientific papers showed, on the one hand, the interest of researchers in the task of assessing the security of big data management*

5 Maria A. Poltavtseva, Dr.Sc. of Technical Sciences, Associate Professor, Peter the Great St. Petersburg Polytechnic University. St. Petersburg, Russia. E-mail: poltavtseva@ibks.spbstu.ru

6 Dmitry P. Zegzha, Corresponding Member of the Russian Academy of Sciences, Dr.Sc. of Technical Sciences, Professor, Federal State Autonomous Educational Institution of Higher Education «Peter the Great St. Petersburg Polytechnic University». St. Petersburg, Russia. E-mail: zegzhda_dp@spbstu.ru

systems, and on the other hand, the lack of estimates proposed for the target class of systems. The authors have formulated security assessment requirements for big data management systems as a specific component of modern information systems. A new security assessment method is also proposed, which for the first time takes into account the specific properties of big data management systems. The proposed method, in addition to the previously proposed estimates, takes into account the disadvantages of access control caused by various data granulation in the target system components. As well, as a large number of trusted users. And, as a result, the need to process confidential data either on trusted nodes or in a hidden (obfuscated or encrypted) form. The proposed estimate is normalized, can be detailed to the evaluation of each specific data processing tool, easily expanded or integrated into higher-level estimates. The reliability and possibility of practical application of the proposed assessment is shown by developing a software prototype based on previously known and tested software solutions.

***Scientific novelty:** the novelty lies in the author's method of assessing the security of big data management systems, which differs for the first time by taking into account the disadvantages of access control caused by different granulation of data and taking into account the trust in individual data processing nodes.*

References

1. Minzov A. S., Nevskij A. Ju., Baronov O. R. Bezopasnost' personal'nyh dannyh: novyj vzgljad na staruju problemu // Voprosy kiberbezopasnosti. 2022. №. 4(50). S. 2–12. DOI:10.21681/2311-3456-2022-4-2-12.
2. Colombo P., Ferrari E. Access control technologies for Big Data management systems: literature review and future trends // Cybersecurity. 2019. T. 2. №. 1. S. 1–13.
3. Rafiq F. et al. Privacy Prevention of Big Data Applications: A Systematic Literature Review // SAGE Open. T.12(2). DOI: 10.1177/21582440221096445
4. Markov A. S., Varenitca V. V., Arustamyan S. S. Topical Issues in the Implementation of Secure Software Development Processes // Proceedings of the International Conference on Information Processes and Systems Development and Quality Assurance IPSQDA-2023 (March 22–24, 2023, St. Petersburg Russia). IEEE. 2023. C. 48–54.
5. Alhazmi H. E., Eassa F. E., Sandokji S. M. Towards Big Data Security Framework by Leveraging Fragmentation and Blockchain Technology // IEEE Access. 2022. T. 10. S. 10768–10782. DOI: 10.1109/ACCESS.2022.3144632.
6. Wang T. et al. Edge-based auditing method for data security in resource-constrained Internet of Things // Journal of Systems Architecture. 2021. T. 114. C. 1–10. DOI: 10.1016/j.sysarc.2020.101971.
7. Stodt, J. at al. Security Audit of a Blockchain-Based Industrial Application Platform // Algorithms. 2021. T. 14(4), 121 c. DOI:10.3390/a14040121.
8. Kalinin M., Poltavtseva M., Zegzhda D. Ensuring the Big Data Traceability in Heterogeneous Data Systems // 2023 International Russian Automation Conference (RusAutoCon). Sochi, Russian Federation. 2023. C. 775–780. DOI: 10.1109/RusAutoCon58002.2023.10272905.
9. Attaallah A. et al. Analyzing the Big Data Security Through a Unified Decision-Making Approach // Intelligent Automation & Soft Computing, 2022. T. 32(2). C. 1071–1088. DOI: 10.32604/iasc.2022.022569.
10. Yang M. Information security risk management model for big data // Advances in Multimedia 2022. T.1 C. 1–10 DOI: 10.1155/2022/3383251
11. Theodorakopoulos L., Theodoropoulou A., Stamatiou Y. A State-of-the-Art Review in Big Data Management Engineering: Real-Life Case Studies, Challenges, and Future Research Directions // Eng. 2024. T. 5(3). C. 1266–1297. DOI: 10.3390/eng5030068.
12. Kalinin M., Poltavtseva M. Big Data Security Evaluation by Bidirectional Analysis of Access Control Policy // 2024 International Russian Smart Industry Conference (SmartIndustryCon). Sochi, Russian Federation. 2024. C. 98–103. DOI: 10.1109/SmartIndustryCon61328.2024.10515459.
13. Poltavtseva, M. A., Zaitseva, V. V., Ivanov, D. V. Assessing the Security of Big Data Systems // Aut. Control Comp. Sci. 2024. T. 58. C. 1352–1364. DOI: 10.3103/S0146411624701025.
14. Dhillon G., Smith K., Dissanayaka I. Information systems security research agenda: Exploring the gap between research and practice // The Journal of Strategic Information Systems. 2021. T. 30. № 4. DOI: 10.1016/j.jsis.2021.101693.
15. Kostogryzov, A. I. Metodicheskie polozhenija po verojatnostnom prognozirovaniju kachestva funkcionirovanija informacionnyh sistem ch. 1–3 / A. I. Kostogryzov, A. A. Nistratov, P. E. Golosov // Voprosy kiberbezopasnosti. 2025. № 2(66). S. 2–19. DOI: 10.21681/2311-3456-2025-2-2-19.
16. Ocenka ujazvimostej avtomatizirovannyh sistem s primeneniem teorii verojatnostej, raspredelenija St'judenta i normal'nyh sluchajnyh velichin / I. V. Atlasov, A. O. Efimov, E. A. Rogozin, A. S. Cherkasova // Voprosy kiberbezopasnosti. 2025. № 2(66). S. 124–131. DOI: 10.21681/2311-3456-2025-2-124-131.
17. Ali, T., Al-Khalidi, M., Al-Zaidi, R. Information Security Risk Assessment Methods in Cloud Computing: Comprehensive Review // Journal of Computer Information Systems. 2024. C. 1–28. DOI: 10.1080/08874417.2024.2329985.
18. Krjukov, R. O., Fedorchenko, E. V., Kotenko, I. V., Novikova, E. S., Zima, V. M. Ocenivanie zashhishhennosti geterogennyh infrastruktur na osnove grafov atak s ispol'zovaniem baz dannyh NVD i MITRE ATT & CK. / R. O Krjukov, E. V Fedorchenko, I. V. Kotenko, E. S Novikova, V. M. Zima / Informacionno-upravljajushhie sistemy. 2024. T. 2., C. 39–50. DOI: 10.31799/1684-8853-2024-2-39-50.
19. Wang J. et al. Big data service architecture: a survey //Journal of Internet Technology. 2020. T. 21. №. 2. S. 393–405.
20. Omotunde H., Ahmed M. A comprehensive review of security measures in database systems: Assessing authentication, access control, and beyond //Mesopotamian Journal of CyberSecurity. 2023. T. 2023. S. 115–133.
21. El Ahdab L. et al. Unified Models and Framework for Querying Distributed Data Across Polystores //International Conference on Research Challenges in Information Science. Cham: Springer Nature Switzerland. 2024. S. 3–18.
22. Wang S. et al. Data privacy and cybersecurity challenges in the digital transformation of the banking sector //Computers & security. 2024. T. 147.
23. Poltavtseva M., Aleksandrova E., Izotova O. Data modeling for consistent access control in heterogeneous big data systems // 2024 Ivannikov Memorial Workshop (IVMEM). Velikiy Novgorod, Russian Federation. 2024. C. 42–48. DOI: 10.1109/IVMEM63006.2024.10659707.
24. Hu V. C. et al. An access control scheme for big data processing //10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing. IEEE. 2014. C. 1–7.