DOI: 10.15514/ISPRAS-2023-35(4)-4

Анализ системы контроля доступа в гетерогенных системах больших данных

М.А. Полтавцева, ORCID: 0000-0001-9659-1244 <*poltavtseva@ibks.spbstu.ru*>
М.О. Калинин, ORCID: 0000-0002-9732-0099 <*max@ibks.spbstu.ru*>

Институт кибербезопасности и защиты информации,
Санкт-Петербургский политехнический университет Петра Великого
195251, Россия, г. Санкт-Петербург, ул. Политехническая, д. 29.

Аннотация. Системы управления большими данными являются сегодня востребованными практически во всех отраслях, они же являются фундаментом для обучения искусственного интеллекта. Использование в системах больших данных гетерогенных полихранилищ привело к тому, что инструменты в рамках одной системы имеют различную грануляцию данных и модели контроля доступа. Согласование таких компонентов администратором безопасности и реализация общей политики доступа сегодня выполняются вручную. Это приводит к росту числа уязвимостей настройки, что, в свою очередь, служит частой причиной утечек данных. Анализ работ в области автоматизации и анализа контроля доступа в системах больших данных показывает отсутствие решений автоматизации для систем на основе полихранилищ. В данной работе ставится задача автоматизации анализа контроля доступа в системах управления большими данными. Авторы формулируют основное противоречие, заключающееся, с одной стороны, в требовании масштабируемости и гибкости контроля доступа, а с другой – в росте нагрузки на администратора безопасности, усугубленное использованием различных моделей данных и контроля доступа в компонентах системы. Для решения этой проблемы предлагается новый автоматизированный метод анализа политик безопасности, основанный на графовой модели обработки данных и позволяющий снизить число возможных уязвимостей, возникающих в результате некорректного администрирования систем big data. При проведении анализа в рамках предложенного метода используется модель жизненного цикла данных в системе, текущие настройки и желаемая политика безопасности. Использование двухпроходного анализа (от источников данных к получателям и обратно) позволяет решить две задачи: анализ системы контроля доступа на возможные уязвимости и проверку соблюдения корректности бизнес – логики. В работе приводится пример анализа политик безопасности системы управления большими данными с использованием разработанного программного прототипа, анализируются полученные результаты.

Ключевые слова: информационная безопасность; большие данные; полихранилища; полибазы данных; контроль доступа; жизненный цикл данных; моделирование обработки данных; политика безопасности.

Для цитирования: Полтавцева М.А., Калинин М.О. Анализ системы контроля доступа в гетерогенных системах больших данных. Труды ИСП РАН, том 35, вып. 4, 2023 г., стр. 93–108. DOI: 10.15514/ISPRAS–2023–35(4)–4.

Благодарности: 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).

Access control system analysis in heterogeneous Big Data management systems

M.A. Poltavtseva ORCID: 0000-0001-9659-1244 <*poltavtseva@ibks.spbstu.ru*>
M.O. Kalinin ORCID: 0000-0002-9732-0099 <*max@ibks.spbstu.ru*>

Institute of Cyber Security and Information Protection,
Peter the Great St. Petersburg Polytechnic University
29, Polytechnicheskaya st., St. Petersburg, 195251, Russia.

Abstract. Big data management systems are in demand today in practically all industries, and they are also the foundation for artificial intelligence training. The use of heterogeneous poly-stores in big data systems has led to the fact that tools within the same system have different data granularity and access control models. Harmonization of such components by the security administrator and implementation of common access-policy is now done manually. This leads to an increasing number of customization vulnerabilities, which in turn serves as a frequent cause of data leaks. Analysis of works in the area of automation and analysis of access control in big data systems shows the lack of automation solutions for poly-store based systems. This paper poses the problem of automating the analysis of access control analysis in big data management systems. The authors formulate the main contradiction, which consists, on the one hand, in the requirement of scalability and flexibility of access control, and on the other hand – in the growth of the burden on the security administrator, aggravated by the use of different data models and access control in the system components. To solve this problem, we propose a new automated method for analyzing security policies based on a graph model of data processing, which reduces the number of possible vulnerabilities resulting from incorrect administration of big data systems. The proposed method uses the data life cycle model of the system, current settings and the desired security policy. The use of two-pass analysis (from data sources to recipients and back) allows to solve two tasks: analyzing the access control system for possible vulnerabilities and checking compliance with correctness of business logic. The paper gives an example of analysis of security policies of the big data management system using the developed software prototype and analyzes the obtained results.

Keywords: information security; big data; polystore; poly-databases; access control; data life cycle; data processing modeling; security policy.

For citation: Poltavtseva M.A., Kalinin M.O. Access control system analysis in heterogeneous Big Data management systems. *Trudy ISP RAN/Proc. ISP RAS*, vol. 35, issue 4, 2023. pp. 93-108 (in Russian). DOI: 10.15514/ISPRAS-2023-35(4)-4.

Acknowledgements. The study was supported by the grant of Russian Science Foundation No.23-11-20003, https://rscf.ru/project/23-11-20003/; grant of St. Petersburg Science Foundation (Agreement No.23-11-20003 on the regional grant).

1. Введение

Понятие больших данных за последние годы стало неотъемлемой частью современной цифровой экономики. Массивы разнородной, динамически меняющейся информации в своей деятельности используют не только поисковые системы и социальные сети, но и торговые интернет – площадки, операторы связи, банковский сектор, сервисы и службы электронного правительства и даже промышленные предприятия.

Экосистема больших данных включает, в общем случае, три архитектурных уровня, на каждом из которых существует свое понимание что такое "большие данные", своя терминология, специалисты, методы и технологии, а также свои угрозы безопасности массивам информации и свои методы защиты.

Эти уровни можно назвать инфраструктурным, логическим (инженерным) и концептуальным (прикладным). На инфраструктурном уровне под большими данными понимаются как правило системы ЦОД – центров обработки данных. В него входят соответствующее оборудование, сетевая инфраструктура, средства виртуализации [1]. На логическом уровне существует термин – системы управления большими данными,

аналогично хорошо известному понятию систем управления базами данных (СУБД). По аналогии с обработкой данных в СУБД этот уровень и называют иногда инженерным – Data engineering. Логический уровень включает в себя новый класс хранилищ данных – полихранилища или полибазы данных, представляющие собой объединение нескольких разнородных СУБД в единую архитектуру обработки информации [2]. Полихранилища также дополняют инструменты потоковой обработки информации и связанные сервисы: брокеры очередей, балансировщики нагрузки и др. На верхнем, концептуальном или прикладном уровне речь идет о ценности данных для бизнеса, организационных кейсах их использования, передаче на аутсорсинг, совместной обработке и других высокоуровневых задачах управления данными и знаниями в организации [3, 4].

Большое число утечек конфиденциальных данных при их обработке в экосистемах больших данных приводит к необходимости ускорения разработки методов и средств обеспечения безопасности для данного класса систем [5, 6]. В данной работе речь пойдет в первую очередь о безопасности логического уровня обработки данных, или безопасности полихранилищ и, в частности, о решении задачи автоматизации анализа контроля доступа.

2. Анализ контроля доступа в системах полихранилищ больших данных

Современные исследования в области повышения эффективности контроля доступа в гетерогенных системах больших данных сконцентрированы в области поиска новых моделей контроля доступа и использования технологий распределенного реестра для защиты от утечек информации.

2.1 Обзор методов средств анализа и анализа контроля доступа в системах полихранилищ больших данных

На сегодняшний день контроль доступа в системах больших данных автоматизирован для инструментов и экосистем, построенных на основе одной модели данных (например, экосистемы Hadoop и модели вида “ключ-значение”). Для систем больших данных, построенных на основе полихранилищ, задача настройки и анализа целостной системы контроля доступа с учетом всех компонентов и грануляции данных в них реализуется вручную. Последние несколько лет исследователями ведется работа в рамках этой задачи по нескольким направлениям.

Во-первых, в области методов и средств контроля доступа на основе одной универсальной модели. Несмотря на использование различных подходов: ролевой модели [7-9], атрибутивной модели [10] и ее модификаций [11], подхода на основе знаний [12] исследователям не удалось преодолеть проблему различной грануляции данных в полихранилищах [13] и проблема несовершенства гранулированного контроля доступа для этого подкласса решений остается открытой.

Только в этом году исследователями в принципе показано, что адаптация современных методов на основе атрибутивного контроля (ABAC) позволяет обеспечить согласованность доступа на уровне системы в целом хотя бы для гомогенной инфраструктуры [14]. Однако предложенное решение подходит только для инфраструктуры Hadoop, основанной на модели данных ключ-значение. При его переносе на другие классы решений и, тем более, распространении на полихранилища уязвимости переноса правил доступа между компонентами с различной грануляцией данных, описанные, в частности, в [15, 13], сохранятся.

Исследователями также предлагаются универсальные средства автоматизации на основе технологии блокчейн [16, 17]. Основная проблема таких средств – также отсутствие аналитического механизма, позволяющего выявить уязвимости реализации контроля

доступа, так как необходимость перехода к внутренним системам разграничения доступа в структурированных хранилищах сохраняется [15, 18].

Таким образом, существующие решения в области автоматизации и анализа контроля доступа в системах больших данных обеспечивают приемлемую безопасность только в рамках гомогенных систем, а в полихранилищах согласование настроек контроля доступа между гетерогенными компонентами остается выполняемым вручную, что приводит к большим временным затратам, высоким требованиям квалификации аналитика безопасности и значимому числу ошибок.

2.2 Проблемы применения методов контроля доступа в системах управления большими данными

Подсистемы контроля доступа в гетерогенных системах управления большими данными сегодня складываются из совместной работы модулей контроля доступа отдельных инструментов обработки информации. В них доминируют традиционные модели доступа, и настройка реализации политик безопасности между гетерогенными компонентами производится целиком вручную. Инструменты автоматизации существуют только в гомогенных решениях, то есть над отдельными сочетаниями компонентов в рамках одного семейства от одного разработчика. Тем не менее, такие особенности больших данных как объем, разнообразие, динамичность во времени приводят к необходимости поиска новых методов контроля доступа в этой области, которые в ближайшей перспективе могут быть интегрированы в промышленные продукты.

Проанализируем анализа современные методы разграничения доступа в системах управления большими данными с точки зрения проблемы автоматизации и анализа контроля доступа в гетерогенных системах больших данных. Ниже приведен ряд методов, предлагаемых различными исследователями или включенных в современные инструменты.

Основным методом контроля доступа в инструментах работы с большими данными сегодня все еще является ролевой контроль доступа (Role-based access control, RBAC) [7] и его модификации [8]. Он же применяется и в промышленных СУБД, интегрированных в системы больших данных на производстве и в корпорациях. Модель RBAC получила широкое распространение среди различных фреймворков (например, Apache Ranger и Apache Sentry [9]). Несмотря на такие достоинства этой модели, как простота и гибкость, ее основным недостатком является плохая масштабируемость и сложность корректного администрирования на большом числе пользователей (ролей).

Вторым направлением в области контроля доступа в системах управления большими данными является атрибутивный контроль доступа (Atribute-based access control, ABAC). Основной единицей данного метода является атрибут – некоторая характеристика объекта, субъекта или среды исполнения. Суть данного метода заключается в написании политик, состоящих из правил сравнения атрибутов субъектов, объектов и др. (например, среды или подключения) [10]. Хорошая масштабируемость, гибкость и универсальность ABAC компенсируется сложностью внедрения и необходимостью отдельной поддержки неструктурированных данных, выраженной в приведенных ниже его расширениях.

Проблему неструктурированных данных стремятся решить авторы таких подходов, как контроль доступа на основе содержимого (Content-Based Access Control, CBAC) [11] и контроль доступа на основе знаний (Knowledge-based access control, KBAC) [12]. Это еще более гибкие методы, чем RBAC и ABAC. Первый использует семантический анализ объектов для принятия решения о доступе, второй – фактически расширяет CBAC автоматизированным процессом нахождения ключевых слов (атрибутов для объектов по мере добавления их в систему и ведением базы знаний.

Несмотря на то, что последние два метода представлены только в теоретических работах, они явно иллюстрируют направление развития систем контроля доступа больших данных в

сторону масштабируемости и, при этом упрощения работы администратора безопасности путем автоматизации. Стоит отметить, что ни один из представленных методов не обладает универсальностью, то есть каждое решение ограничено определенным набором инструментов и предметной областью. В целом обеспечению контроля доступа в системах управления большими данными посвящено немало работ. Помимо упомянутых, отметим работы по интеграции контроля доступа с компонентами шифрования данных с применением современного подхода на основе иммунизации [16] и реализации на основе технологии распределенного реестра (blockchain) [17].

Такое разнообразие говорит о том, что сегодня в системах больших данных стоит ожидать применения различных моделей контроля доступа как на уровне отдельных инструментов, так и на уровне системы в целом. Поэтому на первый план выходит задача согласования этих моделей и политик в автоматизированном (а не ручном, как это проводится сейчас) режиме для минимизации утечек данных и обеспечения корректности бизнес-логики.

Основная проблематика, акцентированная в исследованиях [13, 15, 18], позволяет обозначить общие особенности систем управления большими данными, затрудняющие сегодня внедрение комплексного согласованного контроля доступа на уровне всей системы в целом. Это:

- Большое количество пользователей с доступом различного характера на всем множестве узлов системы, от обслуживающего персонала и администраторов баз данных, до источников данных и их потребителей.
- Сложность и нелинейность жизненного цикла фрагментов информации в системах больших данных, затрудняющая отслеживание фрагментов и согласованное управление доступом.
- Различная структурированность данных в процессе их обработки в системе, в силу использования различных инструментов обработки и множества операций извлечения частей данных и объединения их в новые наборы.

Эти черты приводят к основному противоречию контроля доступа в рассматриваемых системах. С одной стороны, контроль доступа для выполнения бизнес-задач в условиях большого объема и разнородной грануляции данных требует гибкости и масштабируемости, что приводит к большой сложности ее настройки, неточностям и утечкам данных. С другой, при ужесточении контроля доступа на уровне отдельных компонентов или источников данных, снижается вероятность утечек, но такая ситуация может привести к проблемам в реализации бизнес-логики и недоступности части данных для потребителей. Решение данного противоречия заключается в поиске оптимальной (или, по крайней мере, рациональной – если не формулировать задачу как оптимизационную в силу разнородности критериев и оценок для различных организаций) политики безопасности и автоматизации ее выполнения в рамках системы контроля доступа полихранилищ.

Определенную сложность в таком решении представляет собой задача его практической реализации. Как показано выше, при том, что исследователи уже предлагают интеллектуальные системы, основанные на знаниях, сами инструменты еще реализуют RBAC или другие виды контроля доступа, присущие традиционным системам управления базами данных: дистрибутивный, мандатный доступ. Именно этот факт не позволяет применить атрибутивные системы с использованием знаний [14] без дополнительных аналитических компонентов.

Таким образом, в основе контроля доступа в системах управления большими данными должен лежать компонент анализа, позволяющий в автоматизированном режиме выполнить поиск рациональных параметров доступа в конкретной системе с учетом ее компонентов, особенностей работы и бизнес-логики. Отдельным требованием является возможность проведения такого анализа как на этапе проектирования информационной системы,

включающей обработку больших данных, так и уже в процессе ее функционирования, для оценки защищенности и коррекции работы.

3. Метод автоматизированного анализа контроля доступа в гетерогенных системах больших данных

Анализ контроля доступа в системах управления большими данными фактически решает две основные задачи: снижение числа утечек данных из-за ошибок в настройке политик безопасности [19] и обеспечение достаточной информированности потребителей данных для выполнения их бизнес-функций [20]. Для снижения числа утечек данных необходимо проведение оценки всего жизненного цикла данных, так как полученные на более поздних этапах фрагменты в системах Больших данных не только могут быть семантически связаны с исходными, но и содержать их без изменений, а на доступ к исходным данным потребители могут иметь сложные ограничения, обусловленные политикой безопасности [21]. Поэтому при анализе системы контроля доступа требуется как оценка доступа получателей к исходным данным, при заданном доступе к выходной информации, так и оценка доступа получателей к выходным данным, при заданном желаемом доступе к исходным. В результате такого анализа должна быть сформирована общая политика безопасности системы.

Таким образом, в основе автоматизированного анализа контроля доступа находятся:

- Политика доступа к выходным данным, в терминах одной из моделей безопасности (в общем случае – любой). Этот компонент обусловлен бизнес-логикой.
- Желаемая политика доступа к входным данным, в терминах одной из моделей безопасности (в общем случае – любой). Этот компонент обусловлен требованием минимизации доверия.
- Модель обработки данных в системе управления большими данными, представляющую собой описание процесса получения выходного фрагмента из входного.

В модели процесса обработки данных в системе управления большими данными от исходных фрагментов к результирующим должны учитываться права доступа, которые получает каждый производный фрагмент. Для оценки уже функционирующих систем в данном случае могут быть получены настройки доступа напрямую из инструментов обработки автоматическим путем. Решение этой задачи на этапе проектирования систем управления большими данными заключается в формализации и анализе переноса прав доступа в гетерогенном полихранилище [22-24].

Для моделирования процессов обработки данных авторами предлагается использовать графовую модель жизненного цикла данных, предложенную в [25, 26], вершинами которой являются структурированные фрагменты данных, а ребрами – операции над ними. Такой граф описывает процесс обработки больших данных полхранилища и может быть построен в автоматизированном режиме [26], что является значительным преимуществом.

По результатам анализа было принято решение в качестве базового механизма разграничения доступа использовать атрибутивный подход (ABAC), поскольку он обладает достаточной степенью гранулярности и точностью доступа, к тому же, остальные политики могут быть описаны в терминах и правилах ABAC [27, 28]. Надстройка в виде KBAC также может быть приведена к данному виду [29]. Внутри системы анализа исходные правила контроля доступа представляются в терминах следующей атрибутивной модели безопасности:

- A_O, A_S – множество возможных атрибутов объектов и субъектов соответственно;
- $O = \{o_1, o_2, \ldots, o_k\}$ – множество вершин графа обработки информации в системе, представляющие собой фрагменты данных, включая множество $S = \{s_1, s_2, \ldots, s_n\}$ – множество субъектов системы обработки Больших данных.

Для $\forall o_i \in O$, i = 1, …, k, а также для $\forall s_t \in S$, t = 1, …, n определено множество пар «атрибут-значение»: $\{(a_1, v_1), (a_2, v_2), \ldots, (a_m, v_m)\}$, где $a_j \in A_O \cup A_S$, v_j – значение атрибута.

- $D_{вх} \subseteq O$ – множество входных фрагментов данных;
- $D_{вых} \subseteq O$ – множество выходных фрагментов данных;
- E – множество ребер графа обработки информации в системе, представляющие собой операции над фрагментами данных (агрегация и разделение);
- P – политика безопасности – задается как множество правил $\{P_1, P_2, \ldots, P_z\}$, таких что P_i = {conditionsubj, conditionobj, action, access}, где conditionsubj – множество условий для субъекта, conditionobj – множество условий для объекта (фрагмента), action – действие субъекта по отношению к объекту (чтение, запись, чтение и запись), access – разрешение/запрет на выполнение субъектом действия.

Схема метода анализа контроля доступа при анализе на предмет соблюдения бизнес-логики, от входных данных к выходным, приведена на рис. 1.

Проверка графа жизненного цикла данных в процессе обработки проводится на непротиворечивость, наличие ошибок и циклов. Этот шаг необходим в первую очередь для применения анализа на этапе проектирования, когда граф обработки данных строится на основе проектной документации и загружается в систему. Для анализа действующей системы контроля доступа этот шаг с точки зрения безопасности можно опустить, однако он дополнительно позволяет выявить технические проблемы построения процесса обработки данных и может использоваться как технологический инструмент инженером данных.

Далее последовательно рассматриваются все входные фрагменты, и, порожденные ими промежуточные. Последние также добавляются в список для оценки прав, пока не будут получены итоговые, выходные фрагменты данных, не порождающие новых элементов для рассмотрения. Дополнительным преимуществом такого анализа является возможность выявления ошибок в проектировании системы обработки данных, например, наличие неиспользуемых (излишних) данных в промежуточных хранилищах и др.

Аналогичным образом проводится анализ в обратном направлении, от выходных фрагментов данных к входным. Общий порядок проведения обратного анализа таков:

- Формируется список выходных фрагментов данных и устанавливаются права доступа к ним субъектов, на основе настроек системы или предполагаемой плотики безопасности. Формируется список O_{in}.
- Для каждого результирующего фрагмента из списка O_{in} устанавливается, из каких фрагментов данных он сформирован, формируется список O_{temp}. То есть, на основе графа жизненного цикла данных определяются связи по порождению.
- Для каждого фрагмента данных списка O_{temp} устанавливаются возможные права доступа, на основе обратного анализа правил доступа из политики P.
- Элементы O_{temp} переносятся в список O_{in}, отработанные элементы исключаются из рассмотрения (списка O_{in}) и снова запускается шаг 2.
- Формируются итоговые возможные матрицы доступа для входных данных, производится их сравнение с политикой безопасности и настройками контроля доступа.

В результате данный метод анализа позволяет выявить возможные утечки данных, за счет обнаружения доступа потребителей информации и пользователей системы (субъектов) к исходным данным, являющимся для них конфиденциальными.

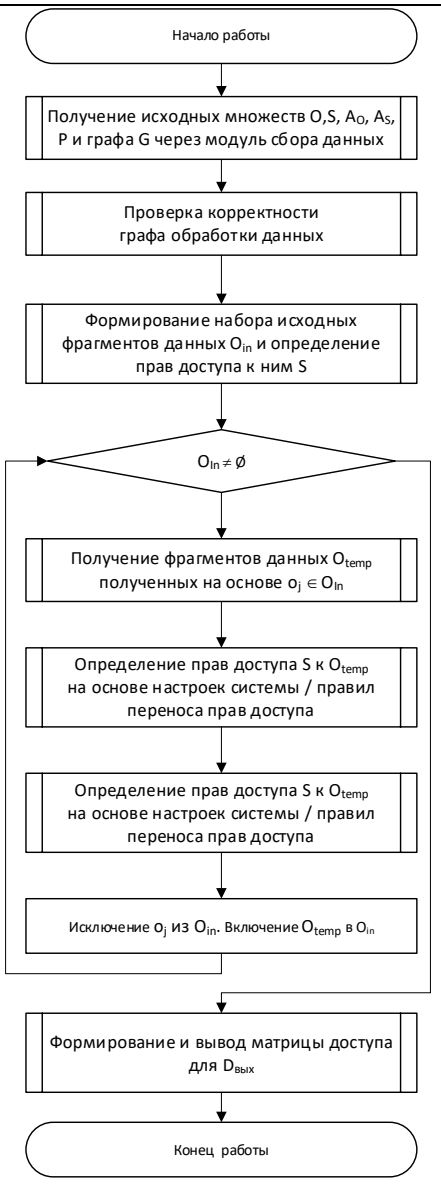

Рис. 1. Метод анализа системы контроля доступа (прямое направление)
Fig. 1. Access control system analysis mechanism (forward direction)

Дополнительно анализ позволяет установить несоответствия требуемым настройкам контроля доступа, и, что является важным, и предложить администратору безопасности возможные варианты иных настроек контроля доступа в инструментах обработки данных полихранилища, отвечающие текущей политике использования выходных данных.

Полученные варианты матриц доступа также могут быть использованы для совершенствования системы.

В то же время, если ни одного удовлетворительного решения не получено, это означает невозможность достичь выполнения требуемой политики безопасности на текущем наборе инструментов с текущим порядком обработки данных. При внесении изменений в эти

компоненты (граф жизненного цикла данных, грануляция и настройка доступа с правилами переноса прав) анализ необходимо повторить.

4. Применение метода анализа системы контроля доступа для гетерогенных систем больших данных

Общая архитектура фреймворка автоматизированного анализа контроля доступа в системах управления большими данными приведена на рис. 2.

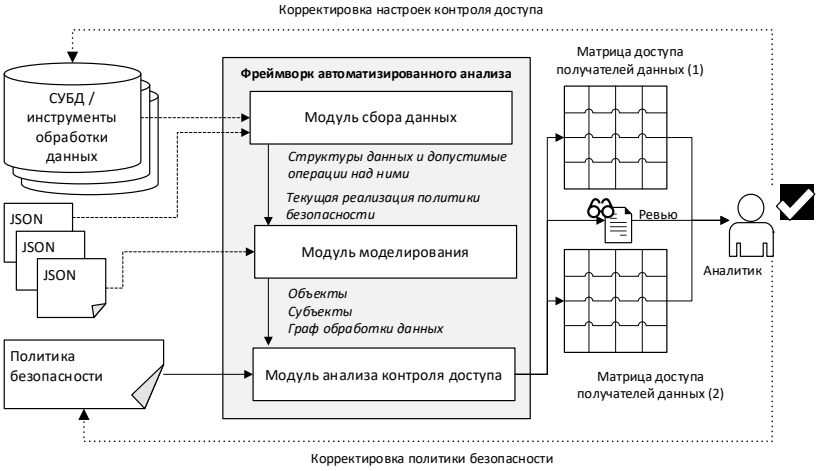

Рис. 2. Архитектура фреймворка автоматизированного анализа контроля доступа в системе больших данных

Fig. 2. Framework architecture for the access control system automation analysis in big data system

Входными данными модуля анализа могут являться как характеристики существующей системы, собранные в автоматическом режиме, так и параметры проектируемой. Выходные данные представляют собой матрицы доступа, полученные на основе имеющейся и желаемой политик безопасности.

Покажем проведение автоматизированного анализа согласно описанному ранее методу на примере. Так как информация в системах больших данных, в том числе та, которая касается контроля доступа и маршрутов обработки данных является конфиденциальной и может быть использована злоумышленниками [30], в статье описан анонимизированный пример системы управления большими данными на основе полихранилища с сокращенным числом пользовательских ролей.

В качестве экспериментального стенда была предложена система обработки данных некоторой организации, состоящей из 4 распределенных филиалов, каждый из которых имеет 4 подразделения: лаборатория, менеджмент, руководство и администратор. Программными компонентами стенда были инструмент потоковой обработки Apache Spark, хранилища данных на основе MongoDB и PostgreSQL. Таким образом, моделировалась гетерогенная среда на с использованием различных моделей данных. Сбор данных о реализованных в СУБД политиках безопасности проводился при помощи интерфейсов запросов и модуля на языке Python. В рамках эксперимента в инструментах обработки использовалась реализованная в них модель RBAC с грануляцией данных в соответствии с возможностями каждого инструмента. Для PostgreSQL – на уровне ячеек (Cell level), для MongoDB – на уровне коллекций (Collection-Level). Отметим, что для MongoDB это наиболее детализованный уровень грануляции доступа. Далее были сформированы таблицы входных фрагментов и субъектов системы вместе с наборами их атрибутов и определена общая

действующая политика безопасности уже в терминах атрибутивной модели. В политику вошли более десяти различных правил разграничения доступа, включая правила на базе предполагаемых настроек доступа конечных пользователей и правила, построенные на основе данных из инструментов обработки. Также сформированы желаемые правила разграничения доступа: желаемые политики безопасности в отношении входных и выходных данных, также в терминах атрибутивной модели. По умолчанию было принято, что если в политике безопасности нет подходящего правила, то доступ запрещен. Кроме того, запрещающие правила находятся в начале политики и имеют приоритет над разрешающими, поэтому если для фрагмента и субъекта на разных этапах жизненного цикла одновременно нашлось запрещающее и разрешающее правила, доступ не предоставляется.

В ходе обработки данных они видоизменяются – объединяются, разделяются и преобразуются. Соответственно, при использовании атрибутивной модели, возникает вопрос наследования не только прав доступа, но и атрибутов в результате этих операций. Если в случае разделения в большинстве случаев достаточно передать атрибуты всем дочерним фрагментам без изменений, то в случае объединения и преобразования все не так просто. В экспериментальной установке в качестве решения для операции объединения предлагается добавлять к атрибуту дополнительную характеристику – тип наследования. Реализованная программа имеет следующие типы наследования атрибутов при порождении новых фрагментов данных:

- max – взятие максимального значения атрибута среди всех родительских;
- min – взятие минимального значения атрибута среди всех родительских;
- concatenate – объединение значений в один список, если они разные.

Правила наследования не относятся к атрибутам субъектов, имеющих доступ к данным, так как наследование прав субъектов в принятой модели безопасности не рассматривается. В итоге была сформирована структура для описания каждого атрибута субъекта или объекта в системе и включения его в решающие правила контроля и переноса доступа. Был сформирован граф жизненного цикла фрагментов данных перехода фрагментов данных (рис. 3). В системе такой граф представлен как матрица смежности с дополнительным хранением весов вершин – типов фрагментов данных и весов ребер – операций над данными.

Пример матрицы доступа, полученной в результате прямого анализа контроля доступа в экспериментальной системе, приведена в табл. 1. Столбцами таблицы являются субъекты системы, а строками – доступ субъектов по отношению к конкретным фрагментам данных. В этой таблице приведены выходные фрагменты 14, 16, 20 и 21.

Полученная в результате анализа матрица доступа прямо указывает на возможное существование нескольких значимых проблем контроля доступа в рассматриваемой системе. Во-первых, если речь идет о полном представлении системы и ее пользователей (а не ее части) отсутствие доступа у всех пользователей к фрагменту 14 говорит о его недоступности для дальнейшего использования. Такой фрагмент (набор данных) должен быть вынесен за пределы системы в архивное хранение или удален. Или же должны быть изменены права доступа кого-либо пользователей в его отношении. Это, очевидно, повлечет фактическое расширение их доступа в отношении входных фрагментов, так как они получат данные на их основе. Такое расширение в свою очередь может нарушить политику безопасности и провести к утечке. Также отметим, что пользователь User 1 не имеет доступа к выходным данным вообще, что для полноценной системы также является не нормальным.

Рассмотрим теперь результат обратного анализа, приведенный в виде примера матрицы доступа из табл. 2. В таблице приведена матрица с минимальными правами доступа, которые должны иметь пользователи к входным фрагментам для получения назначенных им выходных без нарушения политики безопасности.

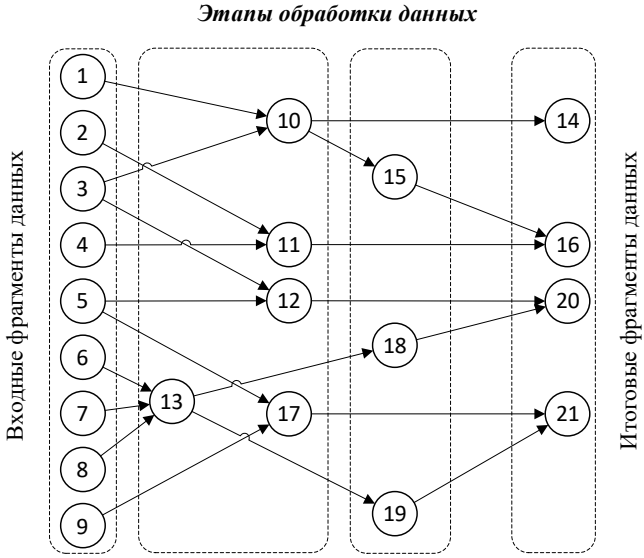

Рис. 3. Граф жизненного цикла фрагментов данных
Fig. 3. Data fragment lifecycle graph

Табл. 1. Матрица доступа для выходных фрагментов (прямой анализ)
Table 1. Access matrix for the output data (forward direction analysis)

ID фрагмента	User 1	User 2	User 3	User 4	User 5
14	Denied	Denied	Denied	Denied	Denied
16	Denied	Read	Read, Write	Read	Denied
20	Denied	Read	Read, Write	Read	Read
21	Denied	Read	Read, Write	Read	Read

Табл. 2. Матрица доступа для входных фрагментов (обратный анализ)
Table 2. Access matrix for the output data (reverse analysis)

ID фрагмента	User 1	User 2	User 3	User 4	User 5
1	Denied	Denied	Denied	Denied	Denied
2	Denied	Read	Denied	Denied	Denied
3	Denied	Denied	Denied	Denied	Denied
4	Denied	Denied	Denied	Read	Denied
5	Denied	Denied	Denied	Denied	Denied
6	Read, Write	Denied	Denied	Denied	Denied
7	Read, Write	Denied	Read, Write	Denied	Denied
8	Denied	Denied	Denied	Denied	Read
9	Denied	Denied	Read, Write	Denied	Denied

Дальнейшей задачей аналитика безопасности является сопоставление полученной матрицы с политикой безопасности организации для обнаружения возможных утечек. Например, действительно ли (User 5) должен иметь доступ на чтение документа 8.

Таким образом, представленный способ автоматизированного анализа контроля доступа в системах управления большими данными позволяет проводить их оценку как на предмет возможных утечек, так и на предмет нарушения бизнес-логики. Анализ может быть проведен на разных этапах жизненного цикла системы, от проектирования до эксплуатации.

5. Анализ полученных результатов

В результате работы авторами выделены основные проблемы применения отдельных методов контроля доступа в системах больших данных. На их основе впервые явно сформулировано основное противоречие между гибкостью и динамичностью настроек контроля доступа в системах больших данных и ростом нагрузки на администратора безопасности, вызванной, в том числе, наличием гетерогенных компонентов полихранилищ в системе обработки информации, не позволяющих эффективно автоматизировать анализ корректности разграничения доступа. Как показано в приведенном обзоре, сегодня эта задача является достаточно новой и даже на уровне исследовательских работ решена только частично: для гомогенных хранилищ экосистемы Hadoop в рамках одной модели контроля доступа. Решения для полихранилищ, кроме приведенного авторами в данной работе, пока не представлено.

Для решения выявленного противоречия авторами предложен новый метод автоматизированного анализа контроля доступа на основе исходной и желаемой политик безопасности и графа обработки данных в системе. Метод включает двухпроходный анализ правил доступа в прямом и обратном направлении над графом обработки данных. Отметим, что результате работы одного цикла разработанного метода для дальнейшей оценки администратору безопасности предоставляются матрицы доступа, позволяющие оценить несколько характеристик системы контроля доступа. При этом

- прямой анализ системы контроля доступа от входных данных к выходным формирует матрицу доступа, позволяющую оценить соблюдение бизнес-логики в системе на основе желаемой политики безопасности;
- обратный анализ от выходных данных к входным позволяет оценить возможность утечек данных в результате логического вывода на основе семантической связности фрагментов данных, проанализировать выполнение принципа минимизации прав доступа.

Итерационное согласование результатов прямого и обратного анализа путем корректировки политик безопасности либо настроек переноса прав доступа отдельных инструментов позволяет администратору (аналитику) безопасности получить рациональную политику контроля доступа для системы в целом, которая может быть также автоматически перенесена в правила конкретных инструментов. Тот факт, что итерационный процесс анализа, за исключением этапа сбора данных, проводится на модели, позволяет снизить нагрузку и избежать излишнего вмешательства в функционирование системы. Отметим, что сбор данных для проведения анализа над действующей системой больших данных, тем не менее, не может проводиться без нагрузки на базовую систему, так как требуется выполнить построение графовой модели обработки информации на основе перемещения фрагментов данных.

Представленный метод и фреймворк автоматизации анализа контроля доступа в системах больших данных в первую очередь позволяет снизить затраты администратора безопасности на сбор информации о контроле доступа в системе больших данных и ручной анализ переноса прав доступа между инструментами обработки. Точный объем времени, сэкономленного в результате автоматизации, к сожалению, не может быть рассчитан, так как на него в значительной степени влияет сложность конкретной системы больших данных и жизненного цикла информации в ней, число типов пользователей системы, число типов разнородных фрагментов данных. Тем не менее, можно говорить о сокращении времени на сбор данных и, главное, на поиск рациональной политики безопасности. Для анонимизированной системы, приведенной в примере данной работы, это время составило несколько часов.

Важно сказать, что представленный метод и фреймворк может быть интегрирован в систему больших данных вместе с другими решениями, обеспечивающими контроль доступа гомогенных компонентов и подсистем обработки информации, описанными ранее в обзоре.

Возможность работы с гетерогенными полихранилищами, в компонентах которых реализованы собственные системы контроля доступа, является его ключевым отличием и преимуществом. В рамках представленного эксперимента использовались СУБД с различной структуризацией данных (пост-реляционная и документоориентированная) относящиеся к разным производителям. Расширение решения на другие классы и конкретные экземпляры систем возможно только путем добавления соответствующих функциональных компонентов к модулю сбора данных, так как организация системных каталогов и диалекты языков запросов в различных решениях существенно отличаются. Также может потребоваться корректировка отображения правил разграничения доступа из инструмента обработки данных в атрибутивную модель. С научной точки зрения некоторую сложность все еще представляет задача построения максимально детализированного графа обработки данных для произвольных хранилищ информации. Ее решение требует интеграции компонентов аудита узлов обработки данных (например, аудита на основе распределенного реестра [26]) с внутренними преобразованиями данных в СУБД при выполнении запросов. Последние могут быть получены на основе анализа планов выполнения запросов реляционных систем и Map-Reduce конвейеров в не реляционных решениях. Сбор таких данных в современных СУБД также хорошо поддается автоматизации, однако требует индивидуальных модулей интеграции с каждым отдельным типом или даже версией системы. «Бесшовное» совмещение таких планов с результатами аудита движения данных между узлами – обработчиками представляется актуальной задачей для дальнейшей работы.

Представленный метод помимо научной новизны обладает следующими преимуществами с точки зрения практической реализации:

- Метод применим для автоматизированного анализа контроля доступа систем управления большими данными на разных стадиях жизненного цикла: как на этапе проектирования, так и на этапе эксплуатации.
- Реализации отельных шагов метода могут быть применены для решения смежных задач, например, анализа качества системы управления большими данными.

Определенными ограничениями применения представленного решения на сегодняшний день является необходимость разработки автоматического модуля сбора данных для каждого инструмента обработки данных, применяемого в целевой системе; все еще относительно высокие требования к квалификации администратора безопасности, и, отчасти, относительно высокая нагрузка прим сборе информации с действующей системы обработки больших данных.

Поэтому дальнейшее развитие представленной работы заключается в повышении степени автоматизации работы путем интеграции с различными типами хранилищ данных, совершенствование алгоритмов и методов сбора и анализа информации. Авторам также представляется рациональной разработка методов поддержки принятия решений при анализе политик безопасности и автоматизация процесса поиска рационального решения на основе методов многокритериальной дискретной оптимизации и алгебры логики.

6. Заключение

Обеспечение безопасности больших данных, даже на одном из уровней рассмотрения, представляет собой сложную и комплексную техническую задачу. Для систем управления большими данными, на уровне обработки информации, она заключается в первую очередь в проблеме преодоления неоднородности и несогласованности программных инструментов, которая усугубляется требованиями бизнес-логики (например, гибкость), масштабом систем и объемов данных, отсутствием стандартов и готовых подходов.

Задача автоматизации анализа контроля доступа является составной частью данной общей проблемы. Получение настроек доступа из инструментов обработки данных и автоматическая их установка в заданные значения – хорошо автоматизированная задача,

тогда как поиск рациональных настроек доступа, в свою очередь, крайне сложен и трудозатратен. Ключевое требование для его реализации – обеспечение согласованности контроля доступа между всеми инструментами, задействованными в процессе обработки данных, в данной работе достигается через модель жизненного цикла фрагментов данных в системе.

Предложенный двухпроходный метод анализа позволяет итерационно достигать рациональных настроек контроля доступа и облегчает проверку соответствия разграничения доступа в реальной системе и требуемой политики безопасности. Анализ в обратном направлении – от получателей к источникам – позволяет установить несоответствия и возможные утечки данных из-за нарушения принципа минимизации привилегий. Анализ в прямом направлении – от источников к получателям в свою очередь дает возможность ускорить поверку соответствия бизнес-логике использования данных.

Предложенный метод и программный прототип средства автоматизации анализа контроля доступа в системах управления большими данными может быть использован в дальнейшем как базис для построения целого ряда научно-технических решений, включая аудит данного класса систем и оценку их защищенности, поиск нарушителя и других.

Список литературы / References

[1]. Mushtaq M. S. et al. Security, integrity, and privacy of cloud computing and big data //Security and Privacy Trends in Cloud Computing and Big Data. – 2022. – pp. 19-51.

[2]. Yung L. R. B., Ströele V., Dantas M. A. R. A Polystore Proposed Environment Supported by an Edge-Fog Infrastructure //International Conference on Advanced Information Networking and Applications. – Cham : Springer International Publishing, 2023. – pp. 292-302. doi: 10.1007/978-3-031-28451-9_26.

[3]. Gao J. Analysis of enterprise financial accounting information manage-ment from the perspective of big data //International Journal of Science and Research (IJSR). – 2022. – vol. 11. – №. 5. – pp. 1272-1276.

[4]. Vasa J., Thakkar A. Deep learning: Differential privacy preservation in the era of big data //Journal of Computer Information Systems. – 2023. – vol. 63. – №. 3. – pp. 608-631. doi: 10.1080/08874417.2022.2089775.

[5]. Dhiman G. et al. Federated learning approach to protect healthcare data over big data scenario //Sustainability. – 2022. – vol. 14. – №. 5. – pp. 1-14. doi: 10.3390/su14052500.

[6]. Strzelecki A., Rizun M. Consumers' Change in Trust and Security after a Personal Data Breach in Online Shopping //Sustainability. – 2022. – vol. 14. – №. 10. – pp. 1-17. doi: 10.3390/su14105866.

[7]. Zhuang Y. et al. Research on big data access control mechanism //International Journal of Computational Science and Engineering. – 2023. – vol. 26. – №. 2. – pp. 192-198. doi: 10.1504/IJCSE.2023.129738.

[8]. Jiang R. et al. T-RBAC Model Based on Two-Dimensional Dynamic Trust Evaluation under Medical Big Data //Wireless Communications and Mobile Computing. – 2021. – vol. 2021. – pp. 1-17. doi: 10.1155/2021/9957214.

[9]. Gupta M., Patwa F., Sandhu R. Object-tagged RBAC model for the Hadoop ecosystem //IFIP Annual Conference on Data and Applications Security and Privacy. – Cham : Springer International Publishing, 2017. – pp. 63-81. doi: 10.1007/978-3-319-61176-1_4.

[10]. Servos D., Osborn S. L. Current research and open problems in attribute-based access control //ACM Computing Surveys (CSUR). – 2017. – vol. 49. – №. 4. – pp. 1-45. doi: 10.1145/3007204.

[11]. Zeng W., Yang Y., Luo B. Content-based access control: Use data content to assist access control for large-scale content-centric databases //2014 IEEE International Conference on Big Data (Big Data). – IEEE, 2014. – pp. 701-710. doi: 10.1109/BigData.2014.7004294.

[12]. El Haourani L., Elkalam A. A., Ouahman A. A. Knowledge Based Access Control a model for security and privacy in the Big Data //Proceedings of the 3rd International Conference on Smart City Applications. – 2018. – pp. 1-8. doi: 10.1145/3286606.3286793.

[13]. Anisetti M. et al. Dynamic and scalable enforcement of access control policies for big data //Proceedings of the 13th International Conference on Management of Digital EcoSystems. – 2021. – pp. 71-78. doi: 10.1145/3444757.3485107.

[14]. Tall A. M., Zou C. C. A Framework for Attribute-Based Access Control in Processing Big Data with Multiple Sensitivities //Applied Sciences. – 2023. – vol. 13. – №. 2. – pp. 1-28. doi: 10.3390/app13021183.

[15]. Colombo P., Ferrari E. Access control technologies for Big Data management systems: literature review and future trends //Cybersecurity. – 2019. – vol. 2. – №. 1. – pp. 1-13. doi: 10.1186/s42400-018-0020-9.
[16]. Muneeshwari P., Athisha G. Extended artificial immune system–based optimized access control for big data on a cloud environment //International Journal of Communication Systems. – 2020. – vol. 33. – №. 13. – p. e3947. pp. 1-15. doi: 10.1002/dac.3947.
[17]. Mounnan O., Abou El Kalam A., El Haourani L. Decentralized access control infrastructure using blockchain for big data //2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA). – IEEE, 2019. – pp. 1-8. doi: 10.1109/AICCSA47632.2019.9035221.
[18]. Vijayalakshmi K., Jayalakshmi V. Shared access control models for big data: a perspective study and analysis //Proceedings of International Conference on Intelligent Computing, Information and Control Systems: ICICCS 2020. – Springer Singapore, 2021. – pp. 397-410. doi: 10.1007/978-981-15-8443-5_33
[19]. 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. – pp. 1-7. doi: 10.4108/icst.collaboratecom.2014.257649.
[20]. Oussous A. et al. Big Data technologies: A survey //Journal of King Saud University – Computer and Information Sciences. – 2018. – vol. 30. – №. 4. – pp. 431-448. doi: 10.1016/j.jksuci.2017.06.001.
[21]. Centonze P. Security and Privacy Frameworks for Access Control Big Data Systems //Computers, Materials & Continua. – 2019. – vol. 59. – №. 2. – pp. 361-374.
[22]. Dziedzic A., Elmore A. J., Stonebraker M. Data transformation and migration in polystores //2016 IEEE High Performance Extreme Computing Conference (HPEC). – IEEE, 2016. – pp. 1-6. doi: 10.1109/HPEC.2016.7761594.
[23]. Kroll J. A., Kohli N., Laskowski P. Privacy and policy in polystores: a data management research agenda //Heterogeneous Data Management, Polystores, and Analytics for Healthcare: VLDB 2019 Workshops, Poly and DMAH, Los Angeles, CA, USA, August 30, 2019, Revised Selected Papers 5. – Springer International Publishing, 2019. – pp. 68-81. doi: 10.1007/978-3-030-33752-0_5.
[24]. Poudel M. et al. Processing analytical queries over polystore system for a large astronomy data repository //Applied Sciences. – 2022. – vol. 12. – №. 5. – pp. 1-23. doi: 10.3390/app12052663.
[25]. Poltavtseva, M. A. Modeling Big Data Management Systems in Information Security / M. A. Poltavtseva, M. O. Kalinin // Automatic Control and Computer Sciences. – 2019. – vol. 53, No. 8. – pp. 895-902. doi: 10.3103/S014641161908025X.
[26]. Poltavtseva M. A. et al. Data protection in heterogeneous big data systems //Journal of Computer Virology and Hacking Techniques. – 2023. – pp. 1-8. doi: 10.1007/s11416-023-00472-3.
[27]. Sahani G., Thaker C., Shah S. Supervised Learning-Based Approach Mining ABAC Rules from Existing RBAC Enabled Systems //EAI Endorsed Transactions on Scalable Information Systems. – 2022. – vol. 10. – №. 1. – pp. 1-8. doi: 10.4108/eetsis.v5i16.1560.
[28]. Talegaon S. et al. Contemporaneous Update and Enforcement of ABAC Policies //Proceedings of the 27th ACM on Symposium on Access Control Models and Technologies. – 2022. – pp. 31-42. doi: 10.1145/3532105.3535021.
[29]. Gupta T., Sural S. Ontology-based Evaluation of ABAC Policies for Inter-Organizational Resource Sharing //Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics. – 2023. – pp. 85-94. doi: 10.1145/3579987.3586572.
[30]. Yang K. et al. An Efficient and Fine-Grained Big Data Access Control Scheme With Privacy-Preserving Policy // IEEE Internet of Things Journal. – vol. 4. – № 2. – p. 563-571. doi: 10.1109/JIOT.2016.2571718.

Информация об авторах / Information about authors

Мария Анатольевна ПОЛТАВЦЕВА – доктор технических наук, доцент, профессор института кибербезопасности и защиты информации, федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого». Область научных интересов: информационная безопасность систем хранения данных и СУБД, безопасность больших данных, сбор, обработка и анализ данных в кибербезопасности, моделирование данных и процессов, мониторинг информационной безопасности крупномасштабных систем.

Maria Anatolyevna POLTAVTSEVA – Dr. Sci. (Tech.), Associate Professor, Professor of the Institute of Cyber Security and Information Protection, Federal State Autonomous Educational Institution of Higher Education "Peter the Great St. Petersburg Polytechnic University". Research

interests: information security of data storage systems and DBMS, big data security, data collection, processing and analysis in cybersecurity, modeling of data and processes, monitoring of information security of large-scale systems.

Максим Олегович КАЛИНИН – доктор технических наук, профессор, профессор института кибербезопасности и защиты информации, федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого». Область научных интересов: информационная безопасность операционных систем, гипервизоров, сред виртуализации и облачных вычислений; методы искусственного интеллекта в кибербезопасности и безопасность интеллектуальных систем

Maxim Olegovych KALININ – Dr. Sci. (Tech.), Professor, Professor of the Institute of Cyber Security and Information Protection, Federal State Autonomous Educational Institution of Higher Education "Peter the Great St. Petersburg Polytechnic University". Research interests: information security of operating systems, hypervisors, virtualization environments and cloud computing; methods of artificial intelligence in cybersecurity and security of intelligent systems.