DOI 10.37882/2223-2966.2024.01.24

ОБЕЗЛИЧИВАНИЕ ПОЛЬЗОВАТЕЛЬСКИХ ДАННЫХ В СИСТЕМАХ БИЗНЕС-АНАЛИТИКИ

DEPERSONALIZATION OF USER DATA IN BUSINESS INTELLIGENCE SYSTEMS

A. Ladikov

Summary. The article considered the types of data processing systems in organizations, typical scenarios for using data in business intelligence systems, the essence and methods of depersonalization of user data, approaches to depersonalization of user data focused on business intelligence systems. Based on the results of the study, a method for depersonalizing personal data in the business intelligence system is proposed, based on a combination of methods recommended by Roskomnadzor and allowing to reduce the limitations of each of the methods considered separately. In the proposed approach, the depersonalization system is endowed with business logic, which obliges it to recognize all incoming data formats and be able to process them. This allows you to increase the level of control over data entering the business intelligence system: unknown data structures and formats will be discarded when processed by the depersonalization system, which reduces the risk of uncontrolled ingress of user data into business intelligence systems. The separation of depersonalization and depersonalization systems allows to increase the degree of protection of users' personal data, while using asymmetric encryption algorithms allows you to store the private key in hardware security systems, which minimizes the risk of decrypting the entire database with personal identifiers when an attacker gets unauthorized access to the depersonalization system.

Keywords: data protection, confidentiality, depersonalization of data, personal data, business analytics.

Ладиков Андрей Владимирович
Руководитель отдела разработки платформы бизнес-аналитики, АО «Лаборатория Касперского»
AndreyNot@mail.ru

Аннотация. В статье были рассмотрены виды систем обработки данных в организациях, типовые сценарии использования данных в системах бизнес-аналитики, сущность и методы обезличивания пользовательских данных, подходы к обезличиванию пользовательских данных, ориентированные на системы бизнес-аналитики. По результатам исследования предложена методика обезличивания персональных данных в системе бизнес-аналитики, построенная на основе комбинации методов, рекомендованных Роскомнадзором, и позволяющая снизить ограничения каждого из методов, рассматриваемых отдельно. В предлагаемом подходе система обезличивания наделяется бизнес-логикой, что обязывает ее распознавать все поступающие форматы данных и уметь их обрабатывать. Это позволяет повысить уровень контроля данных, поступающих в систему бизнес-аналитики: неизвестные структуры и форматы данных будут отброшены при обработке системой обезличивания, что снижает риск неконтролируемого попадания пользовательских данных в системы бизнес-аналитики. Разделение систем обезличивания и деобезличивания позволяет повысить степень защиты персональных данных пользователей, при этом, использование алгоритмов асимметричного шифрования позволяет хранить закрытый ключ в аппаратных системах защиты, что минимизирует риск расшифровки всей базы с персональные идентификаторами при получении злоумышленником несанкционированного доступа к системе обезличивания.

Ключевые слова: защита данных, конфиденциальность, обезличивание данных, персональные данные, бизнес-аналитика.

Введение

Ускоряющаяся в последние годы цифровизация экономики и общества вызывает необходимость обработки большого массива пользовательских данных в организациях, предоставляющих товары и услуги, а также обеспечения конфиденциальности полученной информации. Если при обработке пользовательской информации в бумажном виде её хищение было затруднено необходимостью физического доступа к носителям информации и временем на копирование, то с цифровизацией жизни общества злоумышленники получают новые возможности хищения пользовательских данных. При этом масштабы хищений не ограничиваются конкретным пользователем или кругом лиц. Проникновение в информационную систему позволяет получить информацию о всех пользователях практически с той же легкостью, как информацию об одном конкретном пользователе. Это подтверждается участившимися случаями массовых утечек пользовательских данных, появлением сервисов, занимающихся незаконной коммерциализацией похищенных пользовательских данных: агрегацией результатов различных утечек пользовательских идентификаторов (имя, фамилия, номер телефона, адрес электронной почты, номер паспорта и так далее) [4]. Пользователи информационных систем, соглашаясь на обработку своих данных, ожидают от владельцев информационных систем обеспечения должных мер защиты предоставленной информации, а факт хищения пользовательских данных может нанести как репутационный, так и финансовый ущерб [5, 8].

Системы обработки данных в организациях

Все системы обработки данных в организациях можно разделить на два больших класса систем [10]:

1. Функциональные (оперативные) системы обработки данных, основной целью которых является предоставление пользователю основной функциональности информационной системы, например, открытие банковского счета, запрос документов, электронная оплата платежей и другое.
2. Аналитические системы (системы бизнес-аналитики), данные в которых собираются для целей последующего анализа и повышения эффективности работы информационной системы, организации, упрощения и улучшения пользовательского функционала, законной коммерциализации данных и других аналитических задач.

Ключевым отличием двух данных классов систем с точки зрения безопасности является количество операторов информационной системы внутри организации, имеющих доступ к пользовательским данным. В системах оперативной обработки данных, количество легитимных операторов, имеющих доступ к множеству пользовательских данных, может быть сведено к группе системных администраторов, обслуживающих систему. Разработчики системы ведут работу в тестовой среде с использованием искусственно созданных аналогов пользовательских данных. Доступ легитимных операторов системы, как правило, ограничен единичными запросами к системе. Другие сотрудники организации не имеют доступа к пользовательским данным. В результате подобного разграничения доступа, при условии соблюдения других мер защиты информации существенно снижается риск несанкционированного доступа к системе и, как следствие, риск массовой утечки пользовательских данных. В случае систем бизнес-аналитики количество легитимных операторов системы, имеющих доступ к данным, существенно шире, а в крупных организациях может варьироваться от десятков до сотен инженеров данных, бизнес-аналитиков, специалистов по машинному обучению и других специалистов, которым для выполнения должностных обязанностей необходим доступ к пользовательским данным. Наличие доступа к данным у большого количества специалистов существенно повышает риск утечек данных, добавляя новые вектора атак, такие как появление инсайдеров, атаки на рабочие станции, с которых осуществляется доступ к данным, атаки на пользовательские устройства, которые становятся более актуальными в связи с ростом количества удаленных сотрудников [7, 8]. Стоит отметить, что не все организации разделяют информационные системы на аналитические и системы обработки данных. Однако в данной статье рассматриваются организации, стремящиеся повысить защищенность пользовательских данных, как следствие, стремящихся разделить системы в соответствии с их функциональным назначением.

Типовые сценарии использования данных в системах бизнес-аналитики

Если функциональным системам обработки данных требуется информация по конкретному пользователю, например, поиск пользователя по началу фамилии или первым цифрам номера телефона и извлечение дополнительной информации о пользователях для последующей обработки, то задачи аналитических систем имеют другой характер. Типовыми задачами аналитических систем являются: статистические задачи подсчета общего количества пользователей системы для прогнозирования нагрузки, отнесение пользователя к определенной группе, например, по географическому признаку, по времени использования системы, определение приоритета задач по улучшению функционала информационной системы или приоритета исправления ошибок. Другими примерами аналитических задач могут быть задачи, направленные на повышение эффективности и измеримости бизнеса: определение времени жизни пользователей, их лояльности, среднего чека, наиболее востребованного функционала и других бизнес-задач. Еще одним примером класса задач бизнес-аналитики являются задачи машинного обучения, направленные на выявление похожести пользователей на основе накопленных исторических данных, предсказание потенциальных действий пользователя в системе, например, вероятности покупки того или иного товара или использования той или иной функциональности [8]. Отличительной особенностью приведенных примеров задач является отсутствие необходимости в оригинальных идентификаторах пользователя, что создает возможность использования обезличенных данных.

Обезличивание пользовательских данных

Под пользовательскими данными понимаются любые данные, переданные пользователем в информационную систему. В свою очередь, они могут быть разделены на персональные данные, позволяющие идентифицировать пользователя, и не персональные, не позволяющие идентифицировать конкретного пользователя системы. Очевидно, что обезличиванию подлежат персональные данные.

В статье 3 федерального закона Российской Федерации №152 «О персональных данных» [2] дано следующее определение обезличивания: Обезличивание персональных данных — действия, в результате которых становится невозможным без использования дополнительной информации определить принадлежность персональных данных конкретному субъекту персональных данных. Данное определение схоже с определением псевдоанонимизации, определенной в статье 4 общего регламента по защите персональных данных Европейского Союза (GDPR) [6]. Исходя из данного определения,

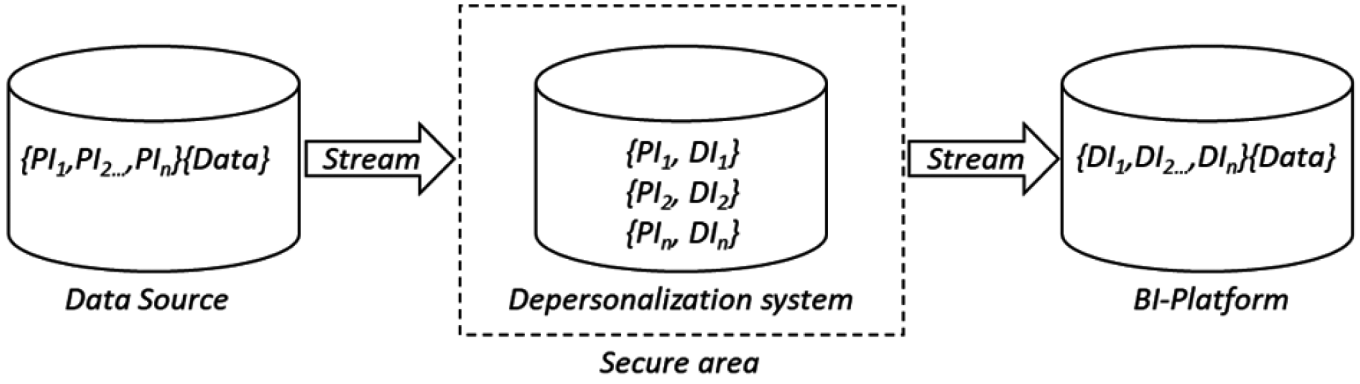
наиболее простым способом обезличивания видится шифрование полей, содержащих персональные данные, сертифицированными алгоритмами шифрования, что соответствует постулату о невозможности определения принадлежности персональных данных конкретному субъекту без дополнительной информации, которой, в данном случае, выступает ключ шифрования. Наиболее существенным ограничением шифрования, как способа обезличивания, является изменение структуры данных — любое семантически значимое поле будет превращено в набор байт произвольного содержания, что потребует существенных изменений структуры хранилищ данных для применения данного метода. В методических рекомендациях Роскомнадзора перечислены наиболее перспективные методы обезличивания персональных данных: метод введения идентификаторов, метод замены состава информации, метод декомпозиции и метод перемешивания [3].

Методика обезличивания пользовательских данных

В основе предлагаемого метода обезличивания лежит метод введения идентификаторов, который в системах электронных платежей называется токенизацией. Суть данного метода заключается в замене персональных данных случайно или псевдослучайно созданными идентификаторами, причем для каждого нового персонального идентификатора случайный идентификатор создается только один раз, при первом его появлении. Таблица соответствия хранится в системе обезличивания. Безопасность данного метода обезличивания обеспечивается реализацией организационно технических методов защиты системы. Схема работы данного метода обезличивания представлена на рисунке 1.

Одним из недостатков данной системы является неполное выполнение свойства релевантности: в случае, если обезличиванию подвергаются данные определенной структуры, например, номер телефона, имя, фамилия или отчество, номер паспорта, они будут заменены численно-буквенными идентификаторами (как пример, будут заменены на GUID). В случае если обезличиванию подвергается уже спроектированная система бизнес-аналитики, этот недостаток существенен, так как типы полей данных и правила их проверки на допустимые значения уже определены. Как следствие, внедрение системы обезличивания, основанной на методе введения идентификаторов, потребует перепроектирования системы бизнес-аналитики, что в ряде случаев требует больших инвестиций. Оптимальным вариантом обезличивания данных в системе бизнес-аналитики является тот, при котором данные сохраняют свою типизацию, и для обезличивания системы достаточно или последовательно произвести обезличивание данных в текущей системе, или подготовить копию используемой системы бизнес-аналитики и наполнить ее данными, предварительно обработав их системой обезличивания. Добиться этого возможно путем выбора случайным образом идентификатора, замещающего персональную информацию из множества допустимых идентификаторов для заданного типа данных. Например, процесс обезличивания телефонного номера может выглядеть следующим образом:

1. Получить на вход системы обезличивания телефонный номер.
2. Определить код страны номера и его формат.
3. Случайно выбрать телефонный номер из множества ранее не использованных номеров заданного формата и пометить номер как используемый (либо сгенерировать номер).



где: *Data Source* - первоисточник пользовательских данных, содержащий персональный данные или другие данные пользователя; *Depersonalization system* – система обезличивания данных, хранящая таблицу соответствия персонального идентификатора и созданного для него обезличенного идентификатора; *BI-Platform* - система бизнес-аналитики; *Secure area* - защищенный периметр системы обезличивания; Stream - потоки данных; P_i – персональный идентификатор; D_i – обезличенный идентификатор, поставленный в соответствие персональному идентификатору P_i; *Data* – неперсональные данные пользователя, ассоциированные с персональными данными.

Рис. 1. Схема работы системы обезличивания методом введения идентификаторов

Источник: составлено автором

4. Сохранить в таблицу соответствия системы обезличивания связь пользовательского телефонного номера и его обезличенного идентификатора.
5. В обрабатываемом наборе данных заменить пользовательский телефонный номер на обезличенный.
6. Сохранить набор данных в систему бизнес-аналитики.

Предлагаемый подход к обезличиванию является комбинацией метода введения идентификаторов и метода перемешивания, с той разницей, что при использовании метода перемешивания допустимым набором данных для замены пользовательского идентификатора на обезличенный является множество поступивших в систему данных. В описанном подходе, обезличенный идентификатор выбирается из всего множества допустимых значений для обезличиваемого типа данных.

Следующим необходимым компонентом системы обезличивания является возможность изменения состава и семантики поступающих данных. К примеру, в случае поступления в систему бизнес-аналитики почтовых адресов или координат местоположения пользователя появляется возможность косвенного его деобезличивания. При этом применить метод перемешивания или метод введения идентификаторов к подобным данным нельзя, так как будет утеряна их аналитическая ценность. Как следствие, для данных, которые не являются однозначно идентифицирующими пользователя, рекомендуется применять другой подход, основанный на изменении состава и семантики данных. Например, в случае обезличивания координат нахождения пользователя может применяться метод их округления.

В предлагаемом подходе система обезличивания наделяется бизнес-логикой, что обязывает ее распознавать все поступающие форматы данных и уметь их обрабатывать. С одной стороны, это приводит к усложнению системы, однако, повышает уровень контроля данных, поступающих в систему бизнес-аналитики: неизвестные структуры и форматы данных будут отброшены при обработке системой обезличивания, что снижает риск неконтролируемого попадания пользовательских данных в системы бизнес-аналитики.

Защита системы обезличивания от утечек

Безопасность пользовательских данных, при использовании метода введения идентификаторов, обеспечивается за счет сохранения в секрете таблицы соответствия персонального идентификатора и сопоставленного ему обезличенного идентификатора, то есть, обеспечивается организационно-техническими методами защиты системы обезличивания. Для понимания способов повышения безопасности таблицы соответствия важно осознавать цели ее использования, которые, могут быть следующими: проверка поступающих пользовательских идентификаторов на наличие в таблице соответствия с целью принятия решения о генерации нового обезличенного идентификатора или использования существующего, путем его извлечения из таблицы соответствия; деобезличивание данных, в случаях, если требуется обработка персональных данных.

Важно отметить, что при использовании метода введения идентификаторов отказаться от таблицы соответствия невозможно, так как она гарантирует работоспособность метода. При этом поддержка сценариев деобезличивания не является обязательным требованием для систем бизнес-аналитики, однако, данный сценарий может быть востребован в случае выявления в данных ошибок или же в случае необходимости деобезличивания данных с целью их передачи в систему, осуществляющую коммуникации с пользователями. Так как первый сценарий является обязательным свойством системы деобезличивания, а второй нет, появляется возможность повышения безопасности системы обезличивания путем выделения системы инструмента деобезличивания данных. Это позволяет отказаться от хранения системы обезличивания персональных идентификаторов в отрытом виде, заменив их, к примеру, на шифрованные идентификаторы. В случае применения асимметричных алгоритмов шифрования система обезличивания может хранить только открытую часть ключа шифрования, что не позволит злоумышленнику сопоставить обезличенные данные с персональными при получении доступа к системе обезличивания [9].

Безопасность метода в этом случае повышается за счет отделения системы обезличивания в отдельную подсистему, которая, в свою очередь, может быть передана на обслуживание отдельному департаменту компании или же третьей, независимой стороне, являющейся гарантом обеспечения безопасности данных в системе (по аналогии с удостоверяющими центрами). Персональные данные пользователей в системе обезличивания являются перемешанными и не несут для злоумышленника практической пользы без получения доступа к системе деобезличивания и системе бизнес-аналитики.

В этом случае алгоритм обезличивания будет следующим:

1. Система обезличивания получает на вход набор данных, требующих обезличивания.
2. Выделяет из полученного набора данные, подлежащие обезличиванию, производя для каждого типа персональных идентификаторов следующую процедуру.

— система обезличивания шифрует персональный идентификатор открытым ключом и производит поиск зашифрованной записи в таблице;

— если зашифрованный персональный идентификатор найден в таблице соответствия, то персональный идентификатор в полученном наборе данных заменяется на найденный в таблице обезличенный идентификатор;
— если зашифрованный персональный идентификатор не был найден в таблице, то для оригинального персонального идентификатора генерируется обезличенный идентификатор; зашифрованный и обезличенный идентификатор сохраняются в таблицу, персональный идентификатор в полученном наборе данных заменяется на созданный обезличенный идентификатор.

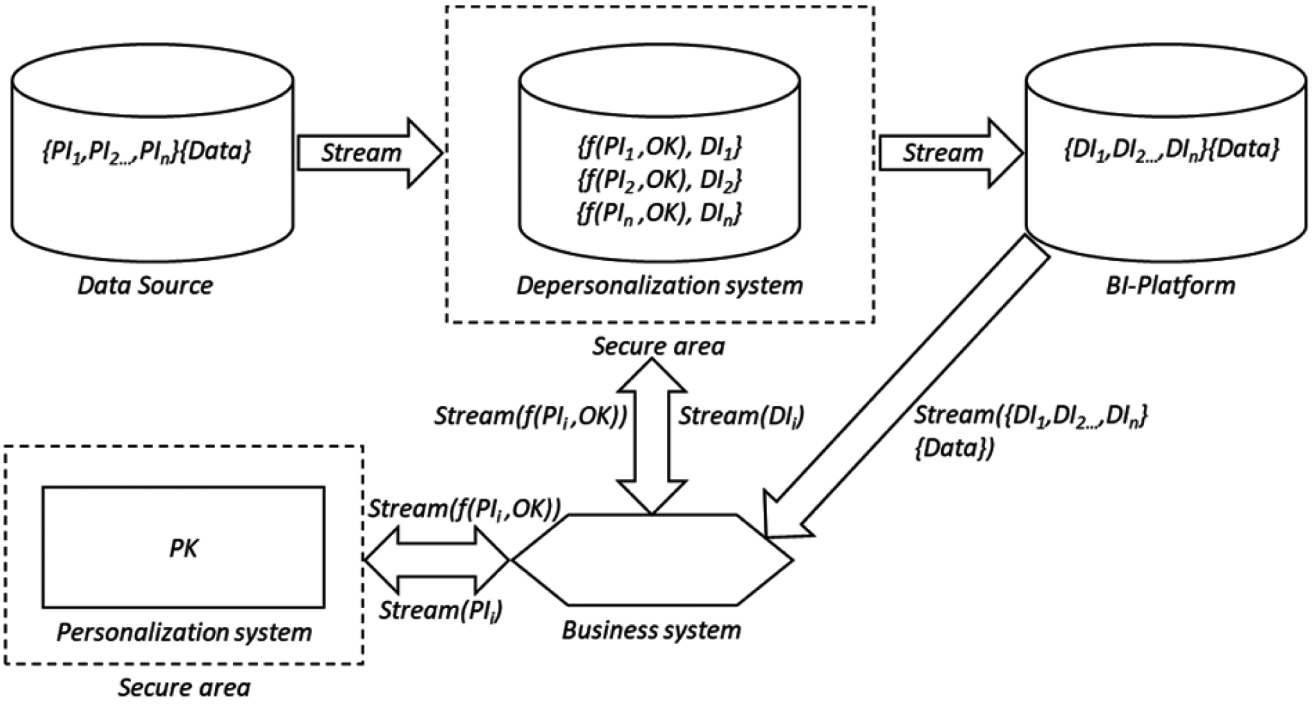
3. Обезличенный набор данных передается в систему бизнес-аналитики для дальнейшей обработки.

Алгоритм деобезличивания будет следующим:

1. Из системы бизнес-аналитики извлекаются данные, содержащие обезличенные идентификаторы.
2. Обезличенные идентификаторы передаются в систему обезличивания, из которой извлекаются шифрованные идентификаторы.
3. Шифрованные персональные идентификаторы передаются в систему деобезличивания.
4. Система деобезличивания производит расшифровку персональных идентификаторов закрытым ключом, после чего передает персональные идентификаторы запрашивающей системе.

Схематически, работа разделенных систем обезличивания и деобезличивания приведена на рисунке 2.

Разделение систем позволяет повысить степень защиты персональных данных пользователей, при этом, использование алгоритмов асимметричного шифрования позволяет хранить закрытый ключ в аппаратных системах защиты, что минимизирует риск расшифровки всей базы с персональные идентификаторами при



где *Data Source* - первоисточник пользовательских данных, содержащий персональный данные или другие идентификаторы пользователя; *Depersonalization system* - система обезличивания данных, хранящая зашифрованные открытым ключом персональные данные и соответствующие им обезличенные идентификаторы; *BI-Platform* - система бизнес-аналитики использующая обезличенные данные; *Personalization system* - система деобезличивания, позволяющая расшифровать персональные данные, зашифрованные в системе обезличивания; *Secure area* - защищенные периметры; *Business system* - система, желающая легитимно деобезличить персональные данные; P_i – персональный идентификатор; D_i – обезличенный идентификатор, поставленный в соответствие персональному идентификатору P_i; *Data* – неперсональные данные пользователя, ассоциированные с персональными данными; $f(PI_1, OK)$ – персональный идентификатор PI_i, зашифрованный открытым ключом *OK* с использованием алгоритма асимметричного шифрования *f()*; *PK* – закрытый ключ, используемый для расшифровки персональных данных системой деперсонализации; *Stream* - потоки данных при обезличивании; $Stream(\{DI_1,DI_2...,DI_n\}\ \{Data\})$ - получение данных из системы бизнес аналитики для обезличивания; $Stream(DI_i)$, $Stream(f(PI_i, OK))$ - запрос по обезличенному идентификатору D_i зашифрованного персонального идентификатора $f(PI_1, OK)$ и возврат данного идентификатора; $Stream(f(PI_i, OK))$, $Stream(PI_i)$ - запрос на расшифровку зашифрованного персонального идентификатора и возврат персонального идентификатора запрашивающей системе.

Рис. 2. Схема деобезличивания данных при разделении систем обезличивания и деобезличивания

Источник: составлено автором

получении злоумышленником несанкционированного доступа к системе обезличивания. Далее приведены практические рекомендации по обезличиванию данных в системах бизнес-аналитики.

1. Для повышения защиты пользовательских данных рекомендуется проводить обезличивание не только данных, относящихся к персональным в соответствии с действующим законодательством, но и любых других данных, которые могут помочь злоумышленнику идентифицировать конкретных пользователей с минимальными усилиями путем сопоставления данных, полученных из скомпрометированной системы, с данными из других источников.
2. При использовании метода введения идентификаторов, рекомендуется формировать релевантные идентификаторы, структура которых соответствует структуре входных данных. Например, при создании идентификатора для номера телефона, идентификатор должен формироваться по формату, соответствующему номеру телефона и состоять из цифр и кода страны, а идентификатор для адресов электронной почты допустимо формировать из английских символов, разделенных символом «@».
3. При обезличивании составных идентификаторов, таких как ФИО, рекомендуется обезличивать каждую его часть отдельно.
4. Данные, не являющиеся идентификаторами, но позволяющие идентифицировать конкретного пользователя по косвенным признакам, например, координаты, IP-адреса, рекомендуется обезличивать методом изменения состава и семантики данных. К примеру, для координат применимо понижение точности, для адресов возможно удаление части адреса.
5. В случае если система бизнес-аналитики предполагает возможность деобезличивания данных, рекомендуется разделить системы обезличивания и деобезличивания. При этом разделение должно быть как техническое, разделяющее системы по различным сегментам сетей и системам, так и организационное, разделяющее администраторов систем.
6. Обработка данных в системе бизнес-аналитики должна выполнять все операции с использованием обезличенных данных, как следствие, система деобезличивания не должна допускать деобезличивания полного набора данных, ограничивая запросы конкретными подмножествами данных, при этом контролируя цели, с которыми оно производится и легитимность субъекта.

Выводы

В статье были рассмотрены виды систем обработки данных в организациях, типовые сценарии использования данных в системах бизнес-аналитики, сущность и методы обезличивания пользовательских данных, подходы к обезличиванию пользовательских данных, ориентированные на системы бизнес-аналитики. По результатам исследования предложена методика обезличивания персональных данных в системе бизнес-аналитики, построенная на основе комбинации методов, рекомендованных Роскомнадзором, и позволяющая снизить ограничения каждого из методов, рассматриваемых отдельно. Предложенные практические рекомендации к обезличиванию данных в системах бизнес-аналитики при незначительной их модификации можно использовать для обезличивания данных в системах с другим функциональным назначением.

ЛИТЕРАТУРА

1. Кодекс Российской Федерации об административных правонарушениях от 30 декабря 2001 г. № 195-ФЗ.
2. Федеральный закон от 27.07.2006 г. №152-ФЗ «О персональных данных» (ред. от 06.02.2023).
3. Приказ Роскомнадзора от 05.09.2013 №996 «Об утверждении требований и методов по обезличиванию персональных данных».
4. Аналитический отчет «Лаборатории Касперского»: «Значимые утечки данных в 2022 году». — URL: https://go.kaspersky.com/ru-data-leakage-report-2022 (дата обращения: 25.12.2023).
5. Виноградова, В.Л., Худякова Н.С., Милованова Л.Р. Защита персональных данных в России: методы и технологии соблюдения регулирования персональных данных // Скиф. — 2023. — №8 (84). — С. 61–65.
6. Общий регламент защиты персональных данных Европейского Союза. — URL: https://gdpr-info.eu/ (дата обращения: 25.12.2023).
7. Селюк, А.С. Защита персональных данных в цифровом пространстве // Вестник Университета имени О.Е. Кутафина. — 2023. — №2. — С. 110–119.
8. Шагапов, И.Р. Защита персональных данных в условиях развития цифровой экономики // Международный журнал гуманитарных и естественных наук. — 2023. — № 1–3 (76). — С. 153–155.
9. Шнайер, Б. Прикладная криптография. Протоколы, алгоритмы и исходный код на С. [пер. с Англ]. — Москва: Диалектика, 2022. — 1040 с.
10. DAMA-DMBOOK: Свод знаний по управлению данными. Второе издание / Dama International [пер. с англ Г. Агафонова]. — Москва: Олимп-Бизнес, 2020. — 828 с.

Журнал «Современная наука: актуальные проблемы теории и практики»