УДК 004.056.5

# МЕТОДЫ ЗАЩИТЫ БОЛЬШИХ ДАННЫХ

***А.В. Ситников***
*Студент 4 курса, кафедра ЭВМ, БГУИР*

***С.Н. Нестеренков***
*Кандидат технических наук, доцент, декан факультета компьютерных систем и сетей*

***А.Н. Марков***
*Старший преподаватель, магистр технических наук, заместитель начальника Центра информатизации и инновационных разработок БГУИР*

*Центр информатизации и инновационных разработок Белорусского государственного университета информатики и радиоэлектроники, Республика Беларусь*
*Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь.*
*E-mail: sitnikov.alexey1@gmail.com, s.nesterenkov@bsuir.by, a.n.markov@bsuir.by*

***А. В. Ситников***
*Студент 4 курса специальности "Вычислительные машины, системы и сети" БГУИР.*

***С. Н. Нестеренков***
*Кандидат технических наук, доцент, декан факультета компьютерных систем и сетей Белорусского государственного университета информатики и радиоэлектроники, доцент кафедры программного обеспечения информационных технологий. Автор публикаций на тему машинного обучения, алгоритмов принятия решений, искусственных нейронных сетей и автоматизации.*

***А. Н. Марков***
*Магистр технических наук, старший преподаватель кафедры ПИКС, заместитель начальника Центра информатизации и инновационных разработок Белорусского государственного университета информатики и радиоэлектроники.*

**Аннотация**. Большие данные – это большой и сложный набор данных, которые невозможно хранить и обрабатывать с помощью традиционного программного обеспечения. Большие данные требуют высокой вычислительной мощности и хранилища, используются распределенные системы. Наборы данных собираются из социальных сетей, медицинских центров, управления данными, учреждений и т.д. Таким образом, конфиденциальность и безопасность данных становятся главной заботой. В этой статье основное внимание уделяется вопросам конфиденциальности и безопасности, а также проблемам конфиденциальности больших данные.

**Ключевые слова**: Big Data, конфиденциальность, безопасность, сохранение конфиденциальности: T-closeness, L-diversity, De-identification.

## Введение.

Большие данные используются во многих приложениях, которые используют Predictive Intelligence, которые люди, проявляют в повседневной жизни [1]. Первым и наиболее распространенным примером является онлайн-реклама, которая предсказывает намерения пользователя при поиске и прочтении документа в Интернете. Компании используют эти данные для обеспечения целенаправленной кампании и привлечения целевой аудитории. Например, если пользователь ищет информацию через Интернет о

покупке камеры, веб-компании могут просмотреть шаблоны поиска и опубликовать объявление о близлежащих магазинах, где можно приобрести камеру и / или наличия скидках в этих магазинах.

**Проблемы конфиденциальности и безопасности в Big Data.**

Конфиденциальность: Конфиденциальность информации – это привилегия иметь некоторый контроль над тем, как собирается и используется личная информация [2-3].

Безопасность: это практика защиты информации и информационных активов с помощью технологий [2-3].

Таблица 1 – Отличительные черты понятий конфиденциальности и безопасности

| Конфиденциальность | Безопасность |
|---|---|
| Конфиденциальность – это надлежащее использование информации пользователей | Безопасность – это “конфиденциальность, целостность и доступность” |
| Конфиденциальность – это возможность решать, какая информация куда поступает. | Безопасность дает возможность быть уверенным в том, что решения будут соблюдаться. |
| Вопрос конфиденциальности часто касается прав потребителей на защиту данных | Безопасность может обеспечить конфиденциальность. Общая цель защищенной системы – защитить предприятие. |

**Требования к конфиденциальности в Big Data.**

Из-за отсутствия стандартных инструментов безопасности и защиты конфиденциальности многие организации решают не пользоваться услугами аналитики больших данных. Разработчики должны иметь возможность проверять соответствие своих приложений с соглашениями о конфиденциальности и конфиденциальной информации независимо от изменений в приложениях и/ или правилах конфиденциальности. На рисунке 1 изображена архитектура больших данных и область тестирования их.

**Конфиденциальность больших данных на этапе генерации данных.**

Генерацию данных можно разделить на активную генерацию данных и пассивную генерацию данных. Методы обеспечения конфиденциальности заключаются в следующем:

1. Ограничение доступа. Если владелец данных считает, что данные могут раскрыть конфиденциальную информацию, которая не должна передаваться, он отказывается предоставлять такие данные.

2. Фальсификация данных.

Под активной генерацией данных подразумевается, что владелец данных предоставит данные третьей стороне, в то время как пассивная генерация данных относится к обстоятельствам, при которых данные создаются в результате онлайн-действий владельца данных (например, просмотра), и владелец данных может не знать о том, что данные собираются третьей стороной.

**Конфиденциальность больших данных на этапе хранения данных.**

Из-за многочисленных технологий, доступных для обработки огромного объема данных, очень сложно хранить данные конфиденциально.

Подходы к сохранению конфиденциальности хранения данных в облаке:

1. Шифрование на основе атрибутов. Контроль доступа основан на идентификации пользователя, чтобы иметь полноценный доступ ко всем ресурсам.

2. Homomorphic шифрование. Может быть развернуто в настройках схемы ABE (шифрование на основе атрибутов, в котором ключ или зашифрованный текст зависят от атрибутов, таких как адрес проживания), возможно обновление приемника зашифрованного текста.

3. Использование гибридных облаков. Гибридное облако – это среда облачных вычислений, которая использует сочетание локальных, частных облачных и сторонних общедоступных облачных сервисов с организацией между двумя платформами.

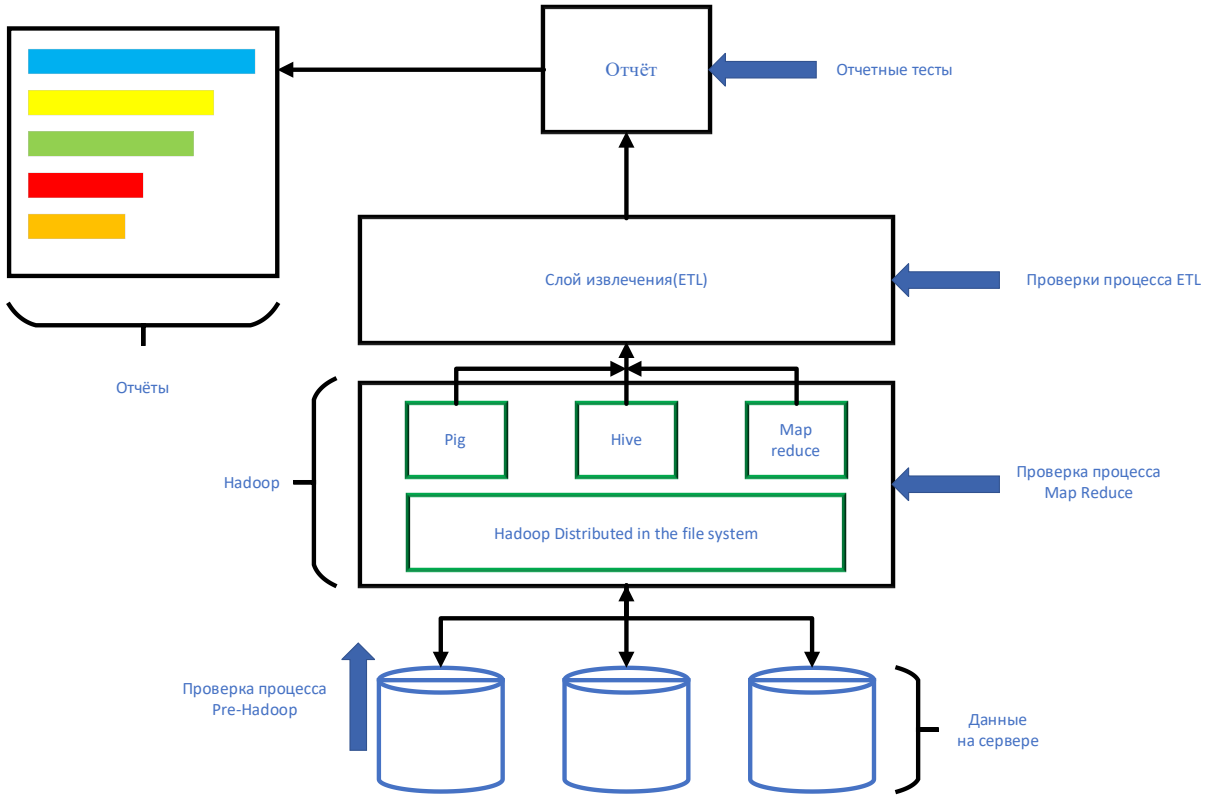


Рисунок 1 – Архитектура больших данных и область тестирования новой парадигмы для тестирования соответствия конфиденциальности в четырех областях процессов ETL (Извлечение, преобразование и загрузка).

**Сохранение конфиденциальности больших данных при обработке данных.**

Для защиты конфиденциальности в части обработки данных можно разделить на два этапа. На первом этапе цель состоит в том, чтобы защитить информацию от несанкционированного раскрытия. На втором этапе цель состоит в том, чтобы извлечь значимую информацию из данных, не нарушая конфиденциальность [4].

**Традиционные методы обеспечения конфиденциальности.**

Деидентификация – это традиционный метод интеллектуального анализа данных с сохранением конфиденциальности. В этом методе данные либо проходят через метод обобщения, либо подавления. При обобщении квазиидентификаторы заменяются более общими, но согласованными значениями, а при подавлении некоторые данные не раскрываются и скрываются с помощью *. Чтобы предотвратить повторную идентификацию данных, были введены понятия k-anonymity, l-diversity и t-closeness [4].

Некоторые общие термины, используемые в полях конфиденциальности этих методов:

1. Атрибут идентификатора – атрибуты, которые однозначно идентифицируют отдельных лиц, например – PAN, no, имя, номер социального страхования и т.д.

2. Квазиидентификатор – атрибут, значение которого в совокупности может однозначно идентифицировать человека, например, пол, возраст, дату рождения и т.д.

3. Конфиденциальный атрибут – информация, которая является частной и личной для отдельного лица, относится к конфиденциальным данным, например, зарплата, болезнь и т.д.

4. Классы эквивалентности – это группа всех записей, которые имеют одинаковое значение в квазиидентификаторах.

**K-anonymity.**

Считается, что таблица обладает k-анонимностью, если каждая запись в таблице похожа по крайней мере на k-1 других записей относительно каждого набора квазиидентификаторов.

Таблица 2 – Неанонимизированная таблица, содержащая записи пациентов

| Имя | Возраст | Почтовый индекс | Заболевание |
|---|---|---|---|
| Владислав | 29 | 47677 | Сердечное заболевание |
| Алексей | 24 | 47602 | Сердечное заболевание |
| Захар | 28 | 47905 | Грипп |
| Рамиль | 27 | 47909 | Грипп |
| Ольга | 24 | 47607 | Грипп |

Существует два обычных метода для завершения K-анонимности при некотором значении k

1. Обобщение. Атрибут "возраст" может быть записан в более общем и широком смысле. Например возраст "19" может быть записан как <= 20.

2. Подавление. В этом методе некоторые значения атрибута скрываются с помощью символа *. Например, некоторые значения почтового индекса могут быть скрыты с помощью *. Таким образом, таблица 3-анонимной версии таблицы 2 будет выглядеть следующим образом(таблица 2):

Таблица 3 – Анонимизированная таблица, содержащая записи пациентов

| Имя | Возраст | Почтовый индекс | Заболевание |
|---|---|---|---|
| Владислав | 20<возраст<=30 | 476** | Сердечное заболевание |
| Алексей | 20<возраст<=30 | 476** | Сердечное заболевание |
| Захар | 20<возраст<=30 | 479** | Грипп |
| Рамиль | 20<возраст<=30 | 479** | Грипп |
| Ольга | 20<возраст<=30 | 476** | Грипп |

Ограничения:

1. K-анонимности недостаточно для предотвращения раскрытия атрибутов.

2. K-аонимность может пострадать от Homogeneity атаки однородности и background knowledge атаки.

**L-diversity.**

Класс эквивалентности считается имеющим L-разнообразие, если для чувствительного атрибута имеются по крайней мере “хорошо представленные” значения (таблица 4).

Таблица 4 – Таблица записи пациентов имеющие L-разнообразие

| Возраст | Зарплата | Заболевание |
|---|---|---|
| 29 | 3k | Грипп |
| 22 | 4k | Гастрит |
| 23 | 5k | Пневмония |
| 52 | 6k | Грипп |
| 43 | 11k | Гастрит |
| 36 | 8k | Пневмония |

Зная, что зарплата пациента находится в диапазоне от 3 до 5 тысяч, то можно сделать вывод, что у него какое-то заболевание, связанное с желудком. Таким образом, происходит

утечка конфиденциальной информации, что приводит к появлению более эффективного метода, называемого t-closeness.

**T-closeness.**

Это дальнейшее усовершенствование L-diversity, основанное на анонимизации, которое используется для сохранения конфиденциальности в наборах данных за счет уменьшения детализации представления данных.

Класс эквивалентности считается имеющим t-близость, если расстояние между распределением чувствительного атрибута в этом классе и распределением атрибута во всей таблице не превышает порогового значения (таблица 5).

Таблица 5 – Таблица записи пациентов имеющие T-closeness

| Возраст | Зарплата | Заболевание |
|---|---|---|
| 2* | 3k | Грипп |
| 2* | 4k | Гастрит |
| 2* | 5k | Пневмония |
| >=40 | 6k | Грипп |
| >=40 | 11k | Гастрит |
| >=40 | 8k | Пневмония |

Ограничения: не касается раскрытия личных данных.

**Сравнение методов**.

Таблица 6 – Вычислительная сложность методов

| Названия метода | Вычислительная сложность |
|---|---|
| K-anonymity | O(klogk) |
| L-diversity | O(n^2/k) |
| T-closeeness | $2^{O(n)O(m)}$ |

**Новейшие методы обеспечения конфиденциальности больших данных.**

Дифференцированная конфиденциальность – один из эффективных методов борьбы с угрозами конфиденциальности [5]. В этой модели личная информация не раскрывается и не изменяется аналитиком для использования. Аналитик не имеет прямого доступа к информации, вместо этого используется посредник, который служит защитой конфиденциальности. Защита конфиденциальности принимает запросы аналитика и выдает результат с небольшими искажениями. Когда риск конфиденциальности невелик, можно рассматривать искажения как неточности, которые достаточно малы, чтобы не влиять на качество ответа, но достаточно велики, чтобы защитить частную жизнь человека.

Шаги по обеспечению дифференциальной конфиденциальности заключаются в следующем (рисунок 2) [5]:

Шаг 1. Аналитик может сделать запрос к базе данных через этого посредника Privacy Guard.

Шаг 2. Защита конфиденциальности принимает запросы и обрабатывает их с помощью базы данных.

Шаг 3. Затем защита конфиденциальности получает ответ из базы данных.

Шаг 4. Защита конфиденциальности добавляет небольшие искажения к данным, а затем, предоставляет аналитику данных.

Преимущества дифференцированной конфиденциальности:

1. Исходные данные не изменяются и не раскрываются конечному пользователю.
2. Нет необходимости в методах обобщения и подавления.
3. Ответ искажается в зависимости от уровня риска, не влияя на качество ответа.

4. Искажение добавляется таким образом, чтобы скрытое значение было полезно аналитику.

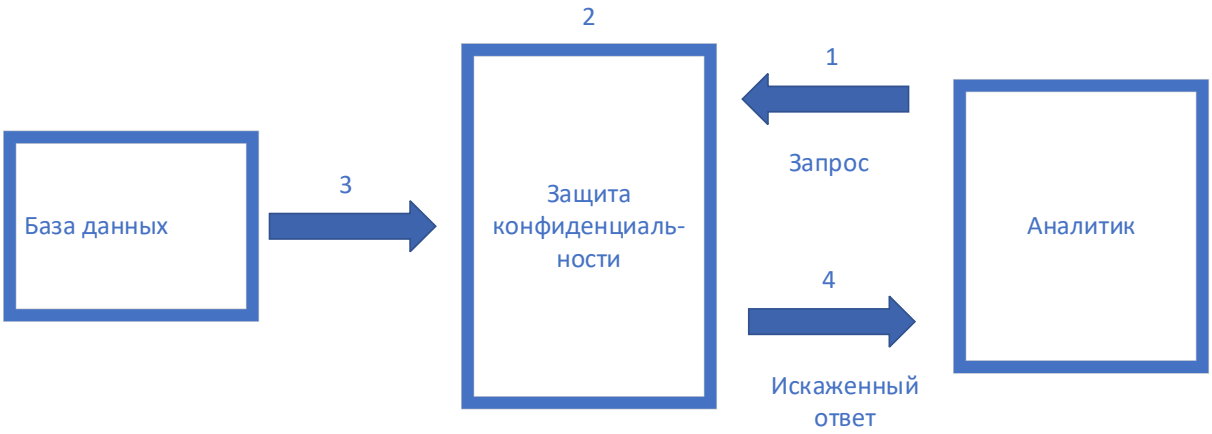


Рисунок 2 – Схема обеспечения дифференциальной конфиденциальности

**Проблемы безопасности конфиденциальности в больших данных.**

1. Повышенный потенциал крупномасштабной кражи или утечки данных
2. Повышенный потенциал крупномасштабной кражи или утечки данных
3. Долгосрочная доступность конфиденциальных наборов данных
4. Проблемы с качеством/целостностью данных и их происхождением
5. Нежелательная корреляция данных и выводы
6. Algorithmic Accountability

**Наиболее существенные риски для конфиденциальности.**

1. Нарушения конфиденциальности и неудобства
2. Анонимизация может стать невозможной
3. Маскировка данных может быть устранена, чтобы раскрыть личную информацию.
4. Неэтичные действия, основанные на интерпретациях.
5. Аналитика больших данных не является точной на 100%.
6. Дискриминация.
7. Для физических лиц существует мало средств правовой защиты
8. Большие данные, вероятно, будут существовать вечно.

**Заключение.**

Большие данные – это большой объем неорганизованных и неструктурированных данных. Конфиденциальность больших данных – очень важный вопрос при организации больших данных. В этой статье были описаны различные методы защиты больших данных в три этапа, то есть генерацию данных, хранение данных и обработку данных, и сделаны выводу о наилучшем методе среди них. Таким образом, в будущем могут быть изучены и внедрены различные методы сохранения конфиденциальности. В области конфиденциальности методов больших данных есть много возможностей в будущем.

***Список литературы***

[1] Нестеренков, С.Н. Применение больших данных в электронном образовании / С.Н. Нестеренков, М.И. Макаров, Н.В. Ющенко, А.Д. Радкевич // BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня : сб. материалов V Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 13-14 марта 2019 года). В 2 ч. Ч. 2 / редкол. : В. А. Богуш [и др.]. - Минск : БГУИР, 2019. - С. 242-245.

[2] Нестеренков, С.Н. Функциональная модель процедур планирования и управления образовательным процессом как основа построения информационной среды учреждения высшего образования / С.Н. Нестеренков, Н.В. Лапицкая // Вести Института современных знаний. - 2018. - N 1. - С. 97-105.

[3] Нестеренков, С.Н. Сетевая модель и алгоритм составления расписания учебных занятий на основе данных прошлых периодов / С.Н. Нестеренков, Н.В. Лапицкая, О.О. Шатилова // Вести Института современных знаний. - 2018. - № 4. - С. 85-92.

[4] Big Data: Concepts, Technology and Architecture / B. Balusamy [и др.]. – Hoboken : John Wiley & Sons, Inc., 2021. – 368 c.

[5] Security and Privacy for Big Data, Cloud Computing and Applications / W. Ren [и др.]. – London : The Institution of Engineering and Technology, 2019. – 328 c.

# BIG DATA PRIVACY METHODS

***A.V. Sitnikov***
*Pregraduate student of the BSUIR*

***S.N. NESTERENKOV,***
*PhD, Associate Professor, Dean of the Faculty of Computer Systems and Networks*

***A.N. MARKOV***
*Senior lecturer of the department, Deputy head of the Center for Informatization and Innovative Developments*

*Center for Informatization and Development of the Belarusian University of State Informatics and Radioelectronics, Republic of Belarus*
*E-mail: sitnikov.alexey1@gmail.com, s.nesterenkov@bsuir.by, a.n.markov@bsuir.by*

**Abstract.** Big data is a large and complex set of data that cannot be stored and processed using traditional software. Big data requires high computing power and storage, distributed systems are used. Data sets are collected from social networks, medical centers, data management, institutions, etc. Thus, privacy and data security become the main concern. This article focuses on privacy and security issues, as well as big data privacy issues.

**Keywords:** Big data, privacy, security, privacy preserving: k-anonymity, T-closeness, L-diversity, De-identification.