

ЗАХИСТ ІНФОРМАЦІЇ В БАЗАХ ДАНИХ ЗА ДОПОМОГОЮ МАСКУВАННЯ

Є. В. Даценко¹

¹ Національний технічний університет України «Київський політехнічний інститут»

Анотація

В даній роботі піднімається проблема витоку персональної інформації у випадку компрометації тестових баз даних. Пропонується альтернативний спосіб вирішення проблеми за допомогою маскування даних та описується альтернативний алгоритм усереднення («averaging»).

Ключові слова: конфіденційність, маскування, база даних, статичне та динамічне маскування.

Вступ

Організації такі, як: банки, страхові агенства, електронна-комерція і т.д. використовують, обробляють та зберігають персональні дані своїх клієнтів. Згідно закону «Про захист персональних даних» [1] захист персональної інформації покладається на власника бази даних. Тому існує ризик компрометації персональних даних у випадку порушення забезпечення комплексної системи захисту інформації. Реальним прикладом є інцидент, який відбувся в лютому 2015 року, коли зловмисники отримали доступ до персональних даних, більше ніж 80 мільйонів користувачів страхової компанії Anthem Inc. [2]. Під персональними даними розуміється інформація, яка однозначно ідентифікує особу. До неї відносять [3]: повне ім'я, домашня адреса, реєстраційний номер облікової картки платника податків, серія та номер паспорта, номер кредитної картки і т.д.

Альтернативним методом захисту персональної інформації висувається вимога по забезпеченню конфіденційності, навіть при отриманні зловмисником доступу до читання бази даних. Методом, який відповідає даним вимогам є маскування. Маскування – це метод створення структурно подібної, але недостовірної версії даних. Проте існуючі алгоритми маскування не забезпечують достатньої подібності маскованих даних до оригінальних. Таким чином, метою роботи була розробка альтернативного методу маскування даних, який забезпечував схожість маскованих даних до оригінальних.

1. Маскування даних

Існує два основних види маскування даних: статичний та динамічний. Статичне маскування це алгоритм при якому дані генеруються шляхом маскування оригінальних даних та зберігаються в окремій базі даних. Динамічне маскування не зберігає замасковані дані, а щоразу генерує їх на основі запиту та прав доступу користувача. Результати порівняльного аналізу видів маскування наведені в таблицях 1 та 2.

Табл. 1. Аналіз статичного маскування

Переваги	Недоліки
В невиробничу БД записуються тільки ті дані, які вже не несуть небезпеки порушення конфіденційності оригінальних даних.	Збій під час транспортування замаскованих даних може спричинити втрату замаскованих даних.
Одноразове навантаження на обчислювальні потужності системи для виконання процесу маскування.	Вимагає додаткового простору для зберігання БД.

Табл. 2. Аналіз динамічного маскування

Переваги	Недоліки
Можливість застосовувати додаткові методи маскування даних.	Існує ймовірність компрометації важливої інформації.
Можливість гнучкого представлення результату запиту, відповідно до прав доступу користувача.	При збільшенні кількості користувачів системи, навантаження на опрацювання запитів буде збільшуватись.

Відповідно до даної таблиці можна зробити висновки, що для забезпечення стійкості маскування метод повинен відповідати наступним вимогам:

- Незворотність. Відсутність можливості відновлення оригінальних даних маючи замасковані;
- Збереження початкового формату даних. Замасковані дані повинні мати форму та вигляд аналогічну до початкових даних;
- Цілісність посилань. Відсутність порушення цілісності посилань між таблицями БД;
- Забезпечення маскування метаданих, що можуть бути використані для відновлення ори-

гінальної інформації, після проведення маскування;

- Збереження особливих властивостей (наприклад, гендерних особливостей).

Існують наступні методи маскування даних, що перетворюють оригінальні дані, в замасковану копію, що не несе корисної інформації для злоумисника. Основними методами вважають наступні:

- Метод заміни – це метод, в процесі якого замінюються оригінальні значення таблиці на значення з спеціального словника.

Недоліком даного методу є обов'язкова наявність словника;

- Метод редагування – це метод, в процесі якого нове значення для заміни створюється заміною всіх літер та цифр універсальним значенням, наприклад, «X».

Переваги: метод можна використовувати як з текстовими так і цифровими даними.

Недоліки: втрачається форма подачі оригінальних даних;

- Метод перемішування подібний заміні, але значення для підміни оригінальних даних беруться з колонки, яка маскується.

Переваги: не потрібно мати словника для заміни.

Недоліки: при малій вибірці даних в таблиці, що потрібно замаскувати, місце оригінальних даних можна отримати перебором – тому є неефективним;

- Метод розмивання – це метод, який замінює значення важливої інформації на випадкове значення в межах певного діапазону;

- Метод усереднення – це метод, який замінює значення таблиці на випадкове значення так, щоб середнє значення вибірки по колонці, яка маскується, залишалося незмінним.

Переваги: дозволяє зберегти середнє значення числових даних до і після маскування.

Недоліки: неможна застосувати для текстових даних.

Проведений аналіз показує, що усереднення є найкращим методом для маскування цифрової інформації. Оскільки забезпечує подібність замаскованих даних до реальних без ймовірності порушення конфіденційності даних.

2. Аналіз існуючих методів усереднення

Метод усереднення використовується у випадку, коли потрібно приховати індивідуальні значення таблиці, при збереженні сукупної картини, наприклад, збереження середнього значення заробітної плати користувачів.

В наведеному прикладі [4] маскується заробітна плата методом усереднення, що працює наступним чином: обчислюється середнє значення заробітної плати з колонки, що необхідно замаскувати, та встановлює це значення кожному працівнику. Приклад виконання даного методу наведений в таблицях 3 та 4.

Табл. 3. Початкові дані бази даних

First name	Salary
David	37500
Layla	55000
Jose	71500
Thomas	95000

Табл. 4. Замасковані дані бази даних

First name	Salary
David	64750
Layla	64750
Jose	64750
Thomas	64750

Метод є швидким, оскільки спочатку обчислює середнє зазначення потрібної колонки, а потім замінює всі оригінальні дані на обчислене значення. На протипагу швидкості є вагомні недоліки, а саме:

- Значення в колонці, що потрібно замаскувати, будуть однакові, що неможливо в реальному житті;
- Після маскування даних значення у всіх однакові, тому ці дані не підходять для тестування нових функцій системи, що пов'язані з фільтрацією даної інформації;
- Якщо сума всіх значень в колонці не буде ділитись націло на кількість елементів, тоді середнє значення до і після маскування не буде співпадати.

Проведений аналіз свідчить про те, що наведений вище метод відповідає вимогам маскування даних, але не забезпечує відповідність маскованих даних до оригінальних, тому метою даної роботи є створення та реалізація нового алгоритму усереднення даних, що виправить наведені вище недоліки.

3. Альтернативний метод усереднення даних

Для виправлення існуючих недоліків методу усереднення пропонується наступний алгоритм, реалізація якого основана на збереженні загальної суми всіх початкових значень, що дасть можливість зберегти їх середнє значення. Функція приймає на вхід назву таблиці та колонки, що потрібно замаскувати. В тілі функції є два цикли: перший – від початку колонки до її середини; другий – від середини до кінця. У першому циклі, проходячи потрібними значеннями, генеруються випадкові числа в діапазоні від 0 до поточного конфіденційного значення та формується нове значення, шляхом віднімання поточного і випадкового значення, де випадкове число записується в окремий масив для подальшого використання. В другому циклі генерується нове значення, шляхом додавання поточного значення таблиці та випадкового значення з масиву, що був створений в попередньому циклі. Після цього, потрібно видалити використане значення з масиву. У випадку непарної кількості елементів в колонці, останнє значення з масиву розділяється на два і записується знову

до масиву, щоб уникнути збереження початкового значення для останнього елементу в колонці.

Для демонстрації власного методу проведено маскування значення заробітної плати в таблиці 5.

Табл. 5. Початкові дані бази даних

ID	Last name	Salary
1	Velasquez	1232
2	Ngao	4492
3	Nagayama	2547
4	Quick-To-See	1247
5	Ropeburn	2331
6	Urguhart	7835
7	Menchu	785
8	Biri	1472
9	Catchpole	759

Після проведення маскування отримали таблицю 6.

Табл. 6. Замасковані дані бази даних

ID	Last name	Salary
1	Velasquez	405
2	Ngao	2746
3	Nagayama	1772
4	Quick-To-See	335
5	Ropeburn	3243
6	Urguhart	8662
7	Menchu	1560
8	Biri	2345
9	Catchpole	1632

Проаналізувавши таблиці 5 та 6, отримали наступні результати:

- Середнє значення колонки заробітної плати становить 2522.22 в обох таблицях;

- Замасковані дані неоднакові.

Тому наведений алгоритм кращий за попередній, оскільки корегує недоліки існуючих алгоритмів.

Висновки

Зараз банки, онлайн-сервіси та інші організації, що зберігають персональну інформацію про своїх клієнтів, мають забезпечувати конфіденційність цієї інформації. Використовуючи маскування даних, можна забезпечити конфіденційність інформації навіть при компрометації баз даних та отриманні зловмишником доступу до інформації з обмеженим доступом.

Результатом роботи стало створення та реалізація нового методу усереднення даних, в якому були враховані та виправлені недоліки існуючих методів.

Перелік використаних джерел

1. Закон України «Про захист персональних даних» від 01.06.2010 №2297-VI (редакція станом на 21.05.2016). — 2010. — URL: <http://zakon.rada.gov.ua/go/2297-17>.
2. Data Breach at Health Insurer Anthem Could Impact Millions. — 2015. — URL: <http://krebsonsecurity.com/2015/02/data-breach-at-health-could-impact-millions/>.
3. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII). — 2010. — P. 59 c. — URL: <http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf>.
4. Overview of Data Masking Methods. — 2014. — URL: <http://smartbridge.com/overview-data-masking-methods/>.