Bezpieczeństwo i koszt ochrony statystycznych baz danych wykorzystywanych w jednostkach administracji publicznej

Łukasz Ślęzak
Jarosław Butanowicz

Wojskowa Akademia Techniczna im. Jarosława Dąbrowskiego

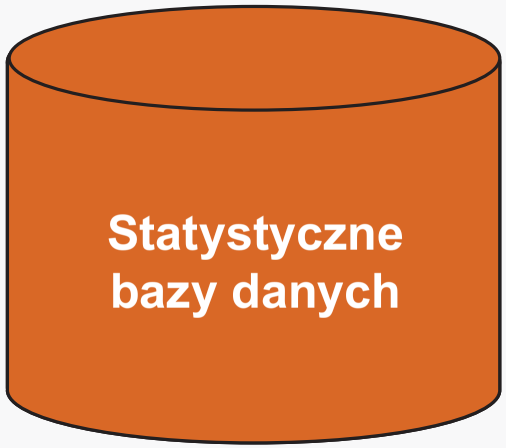
Zaufanie i bezpieczeństwo

- Jednym z kluczowych obszarów działań Europejskiej agendy cyfrowej jest **Zaufanie i bezpieczeństwo.**

  Europejczycy nie będą korzystać z technologii, którym nie ufają. Epoka cyfrowa to nie „wielki brat" ani „cybernetyczny dziki zachód".

- Bezpieczeństwo informacji jako czynnik:
  - nieograniczający prowadzenia działalności operacyjnej,
  - umożliwiający prowadzenie działalności operacyjnej,
  - stwarzający możliwość rozwoju działalności operacyjnej.

Problem ochrony baz danych

Bazy danych		
Problem ochrony	- Identyfikacja i uwierzytelnienie użytkownika	- Udostępnianie statystyk obiektów, przy jednoczesnym ograniczeniu możliwości dostępu do informacji o indywidualnym obiekcie - Korelacja statystyk - Wiedza dodatkowa użytkownika

Statystyczne bazy danych – przykłady zastosowań

Przykłady zastosowań statystycznych baz danych:	
Ministerstwo Zdrowia	**Informatyzacja ochrony zdrowia** ► Wdrożenie Elektronicznej Platformy Gromadzenia, Analizy i Udostępniania zasobów cyfrowych o Zdarzeniach Medycznych – Zbiory (bazy) danych medycznych.
GUS	**Spisy Powszechne** ► Spis próbny do Narodowego Spisu Powszechnego Ludności i Mieszkań (1 kwietnia – 31 maja 2010). ► Powszechny Spis Rolny 2010 (1 września – 31 października 2010). ► Narodowy Spis Powszechny Ludności i Mieszkań 2011 (1 kwietnia – 30 czerwca 2011).

Przykład – Uczestnicy Forum TI 2011

Imię i Nazwisko	Miejsce urodzenia	Firma	Zarobki miesięczne
Adam Nowak	Warszawa	DBSA	900 PLN
Adam Wójcik	Warszawa	HackerCorp	1100 PLN
Aneta Kowalczyk	Warszawa	Computer	1200 PLN
Arkadiusz Malinowski	Warszawa	HackerCorp	1240 PLN
Jarosław Butanowicz	Jastrzębie Zdrój	Sabre	1000 PLN
Konrad Dąbrowski	Warszawa	HackerCorp	680 PLN
Łukasz Ślęzak	Warszawa	Ernst & Young	1000 PLN
Mirosław Kowalski	Warszawa	Uni	990 PLN

Zanonimizowane dane

Ujawnianie statystyk wrażliwych – wybrane przykłady (1/5)

- Statystyka wrażliwa – statystyka, która ujawnia informacje, które mogą doprowadzić do ujawnienia informacji dotyczących indywidualnego obiektu.
- Statystyka obliczana na podstawie poufnych informacji ze zbioru odpowiedzi o liczności 1 jest zawsze wrażliwa.
- Przykład:
  - Wiemy, że miejsce urodzenia Jarosława Butanowicza to Jastrzębie Zdrój.
  - Ilość (Miejsce urodzenia = Jastrzębie Zdrój) = 1.
  - Suma (Zarobki miesięczne dla Miejsce urodzenia = Jastrzębie Zdrój) = 1000 PLN.
  - **Wniosek: Zarobki miesięczne Jarosława Butanowicza są równe 1000 PLN.**
- Są to najprostsze sposoby ujawniania statystyk wrażliwych tzw. ataki z użyciem małych lub dużych zbiorów odpowiedzi.

Ujawnianie statystyk wrażliwych – wybrane przykłady (2/5)

- Przykład ataku z użyciem dużego zbioru odpowiedzi:

  - Analogicznie do poprzedniego przykładu jeśli wiemy, że miejsce urodzenia wszystkich uczestników konferencji poza Jarosławem Butanowiczem to Warszawa.
  - Ilość (Wszyscy) = 8.
  - Ilość (Miejsce urodzenia = Warszawa) = 7.
  - Suma (Zarobki miesięczne dla Wszyscy) = 8110 PLN.
  - Suma (Zarobki miesięczne dla Miejsce urodzenia = Warszawa) = 7110 PLN.
  - **Wniosek: Zarobki miesięczne Jarosława Butanowicza są równe:**
    Suma (Zarobki miesięczne dla Wszyscy) - Suma (Zarobki miesięczne dla Miejsce urodzenia = Warszawa) = **1000 PLN.**

Ujawnianie statystyk wrażliwych – wybrane przykłady (3/5)

- Ochroną przed atakami z użyciem małych lub dużych zbiorów odpowiedzi jest technika oparta na sterowaniu licznością zbioru odpowiedzi:

  Odrzucenie zapytania odwołującego się do zbioru odpowiedzi mających mniej niż *n* lub więcej niż *N* – *n* rekordów,

  gdzie:

  N – liczba wszystkich rekordów,

  n – dodatnia liczba całkowita mniejsza niż *N*.

- Technika sterowania licznością zbioru odpowiedzi może być rozszerzona o wskaźnik dominancji, czyli dodatkowo nieujawniane są zbiory odpowiedzi, w których suma wartości rekordów jest większa od *k%* sumy całej populacji.
- Przykład:

  Suma dochodów dużej firmy oraz małego sklepu.

Ujawnianie statystyk wrażliwych – wybrane przykłady (4/5)

- Sterowanie licznością zbioru odpowiedzi nie jest rozwiązaniem gwarantującym bezpieczeństwo statystycznych baz danych.
- Przykładem łatwego do realizacji ataku na statystyczną bazę danych zabezpieczoną powyższą techniką jest użycie prostej techniki penetracji zwanej szperaczem.
- Wartym podkreślenia jest fakt, iż przeniknięcie jest możliwe nawet dla *n* bliskiego *N/2*.
- Podstawową ideą tej techniki jest:
  - uzupełnianie małych zbiorów odpowiedzi taką liczbą dodatkowych rekordów, aby było możliwe uzyskanie odpowiedzi,
  - odjęcie odpowiedzi opartej na rekordach dodatkowych.

Ujawnianie statystyk wrażliwych – wybrane przykłady (5/5)

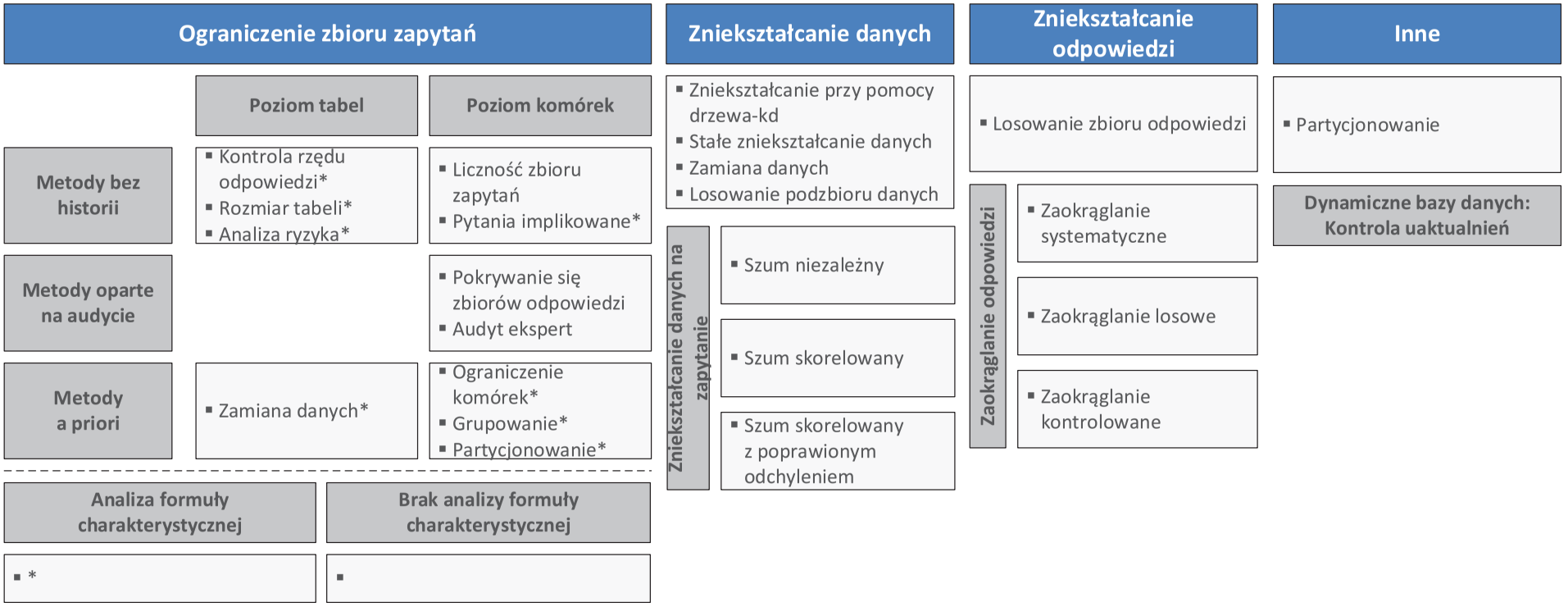
- Przykład z wykorzystaniem szperacza indywidualnego:
  - Szukamy zapytania dozwolonego:
    Suma(Zarobki dla Firma = HackerCorp) = 3020 PLN.
  - Zapytanie odwrotne: Suma(Zarobki dla Firma != HackerComp) = 5090 PLN.
  - A = Suma(Zarobki dla Wszyscy) = 3020 + 5090 = 8110 PLN.
  - B = Suma(Zarobki dla (Firma = HackerComp lub Miejsce urodzenia = Jastrzębie Zdrój)) = 4020 PLN.
  - C = Suma(Zarobki dla (Firma != HackerComp lub Miejsce urodzenia = Jastrzębie Zdrój)) = 5090 PLN.
  - **Wniosek: Wykorzystując szperacz możemy stwierdzić, że Zarobki miesięczne Jarosława Butanowicza są równe**:
    X = B + C – A = 4020 + 5090 – 8110 = **1000 PLN.**

Techniki ochrony statystycznych baz danych

Obecnie istnieje wiele mechanizmów ochrony statystycznych baz danych (przy zachowaniu statystycznej poprawności uzyskiwanych statystyk), które możemy podzielić na cztery główne grupy:	
Ograniczenie zbioru zapytań	Wprowadzenie ograniczeń dostępu do statystyk na podstawie rodzaju zapytania.
Zniekształcanie danych	Wprowadzenie szumów wyłącznie w odpowiedziach na zapytanie o statystykę, bez modyfikacji samych danych.
Zniekształcanie odpowiedzi	Wprowadzenie pewnej permutacji w danych przechowywanych w bazie danych.
Inne	Pozostałe techniki.

Techniki ochrony statystycznych baz danych

- Przykłady technik:



Parametry ochrony statystycznych baz danych (1/3)

- Techniki ochrony statystycznych baz danych zniekształcające dane pierwotne lub odpowiedzi, co wpływa na wartość otrzymywanych statystyk.
- Oprócz zależności pomiędzy bezpieczeństwem danych a dokładnością otrzymywanych statystyk, ważnym elementem jest koszt implementacji danej techniki, a także jej wpływ na efektywność działania bazy danych.
- Dla celów analizy porównawczej istniejących technik względem siebie została wyodrębniona podstawowa grupa parametrów ochrony statystycznych baz danych.

Parametry ochrony statystycznych baz danych (2/3)

Parametr	Definicja
Bezpieczeństwo	Odporność bazy danych na ataki użytkownika nie posiadającego dodatkowej wiedzy.
Utrata zdolności informacyjnych	Mierzona przez liczbę niewrażliwych statystyk, które są zastrzeżone dla użytkownika w związku z mechanizmami bezpieczeństwa.
Koszt	Koszt implementacji jak i koszt przetwarzania zapytań.
Precyzja*	Odchylenie zniekształconych wyników zapytań od rzeczywistych odpowiedzi.
Konsekwencja*	Zdolność systemu bazy danych do odpowiedzi na zapytania z uniknięciem błędów logicznych.

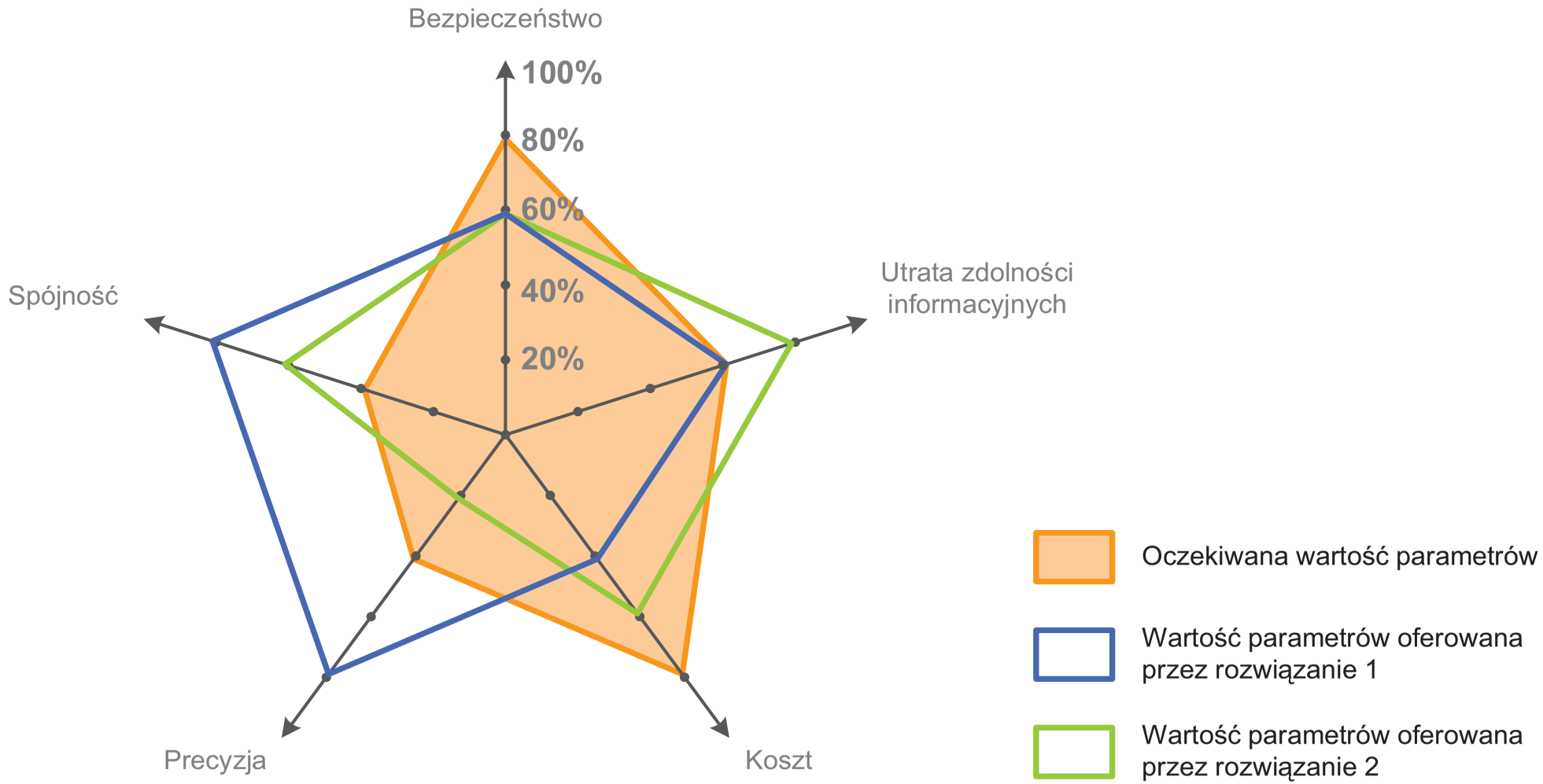
* Precyzja i Konsekwencja są parametrami stosowanymi wyłącznie do ewaluacji technik zniekształcania danych i odpowiedzi.

Parametry ochrony statystycznych baz danych (3/3)

- Koszt

- Parametr, który przedstawia koszt implementacji mechanizmu ochrony, a także koszt dodatkowego procesowania każdego zapytania bazodanowego.
- Metoda pomiarowa sprawdzając dodatkowy koszt implementacji wyrażona jest w pieniądzach.
- Natomiast dodatkowy koszt procesowania zapytań bazodanowych wyrażony jest poprzez różnicę w czasie procesowania danego zapytania dla systemu bazy danych z wdrożonym mechanizmem ochrony bądź bez.
- Znaczenie każdego z tych dwóch czynników jest określane indywidualnie dla każdej organizacji.

Porównanie mechanizmów ochrony (1/2)



Porównanie mechanizmów ochrony (2/2)

- Porównanie mechanizmów ochrony statystycznych baz danych pod kątem poszczególnych parametrów ochrony statystycznych baz danych:

Technika	Bezpieczeństwo	Utrata zdolności informacyjnych	Koszt	Precyzja	Konsekwencja
Stałe zniekształcenie danych	wysokie	wysokie	**średni**	słaba	wysoka
Zamiana danych	średnie	wysokie	**wysoki**	średnia	wysoka
Losowanie podzbioru danych	średnie	średnie	**niski**	średnia	wysoka
Zniekształcenie danych na zapytanie	średnie	średnie	**niski**	średnia	słaba
Zaokrąglanie odpowiedzi	średnie	średnie	**niski**	średnia	wysoka

Q & A

- Łukasz Ślęzak

  lslezak@wat.edu.pl

Bibliografia

- Dorothy Elizabeth Robling Denning, Cryptography and Data Security, 1982.
- G. T. Duncan and S. Mukherjee. Optimal disclosure limitation strategy in statistical databases: deterring tracker attackes through additive noise. *Journal of the American Statistical Association*, 95:720--729, 2000.
- SCHATZ J. M. Survey of Techniques for Securing Statistical Databases
- MURALIDHAR, K AND SARATHY, R. 1999. Security of Random Data Perturbation Methods - *ACM Transactions on Database Systems*, Vol. 24, No. 4, December 1999.
- ADAM, N. R. AND WORTMANN, J. C. 1989. Security-control methods for statistical databases: A comparative study. *ACM Comput. Surv. 21*, 4 (Dec. 1989), 515–556.
- TENDICK, P. AND MATLOFF, N. 1994. A modified random perturbation method for database security. *ACM Trans. Database Syst. 19*, 1 (Mar. 1994), 47–63.
- REISS, S. P. 1983. Practical Data-Swapping: The First Steps. *ACM Transactions on Database Systems*, Vol. 9, No. 1, March 1984.