WIADOMOŚCI STATYSTYCZNE

PL ISSN 0043-518X
e-ISSN 2543-8476

Rok LXII
3 (670) 2017, 71—85

Bolesław SZAFRAŃSKI

Podstawy budowy skutecznych metod ochrony statystycznych baz danych

Streszczenie. *Głównym celem artykułu jest zwrócenie uwagi na znaczenie prac badawczo-rozwojowych dotyczących metod ochrony statystycznych baz danych w sytuacji, gdy uniwersalne systemy zarządzania bazami danych nie mają mechanizmów wspierających w wymaganym stopniu bezpieczeństwa baz statystycznych. W pracy przedstawiono podstawowe modele sterowania dostępem i przepływem danych oraz wykazano ich ograniczoną przydatność dla statystycznych baz danych. Omówiono też specyfikę problemu bezpieczeństwa tego typu baz danych oraz metod ataku na nie i ich ochrony.*

W konkluzji stwierdzono, że w Polsce nie obowiązują powszechnie uznane, teoretyczne podstawy budowy bezpiecznych mechanizmów ochrony statystycznych baz danych. Prace dotyczące ochrony danych mają charakter przyczynkarski, odległy od możliwości komercyjnego wdrożenia uzyskanych wyników. Istnieje zatem potrzeba prowadzenia prac interdyscyplinarnych GUS i uczelnianych zespołów badawczych.

Słowa kluczowe: baza danych, ochrona danych, sterowanie: dostępem, przepływem i wnioskowaniem, statystyczna baza danych.

JEL: C80, C81

Od wielu lat obserwujemy wzrost znaczenia analizy statystycznej lub eksploracji danych i jednocześnie brak zaawansowanych prac badawczo-rozwojowych uwzględniających specyfikę badań statystycznych, skutkujących wzmocnieniem mechanizmów ochrony danych w dostępnych komercyjnie systemach zarządzania bazami danych. Dlatego głównym celem niniejszego artykułu jest zwrócenie uwagi na tę zasadniczą dysproporcję, zwłaszcza w sytuacji gdy statystyka publiczna w każdym kraju jest jednym z podstawowych systemów informacyjnych

zapewniających niezbędne bezpieczeństwo informacyjne[1] społeczeństwa, gospodarki i państwa. Taką właśnie rolę pełni system informacyjny GUS jako nieodłączny element systemu informacyjnego Polski. Na marginesie warto podkreślić, że ważną jego funkcją, oprócz dostarczania informacji i wiedzy statystycznej, jest porządkowanie ogólnokrajowej infrastruktury informacyjnej dzięki rejestrom klasyfikacyjnym wykorzystywanym w badaniach statystycznych. W przypadku braku takiego uporządkowania (lub inaczej, braku „narzucanego" przez nie ładu informacyjnego) nie jest możliwe prowadzenie niezbędnych, z punktu widzenia zarządzania państwem, analiz porównawczych w układzie wewnętrznym (krajowym) i zewnętrznym (międzynarodowym).

Podstawą zmian unowocześniających działalność GUS są cele i inicjatywy strategiczne zawarte w dokumencie *Kierunki rozwoju statystyki publicznej do roku 2017*. Dla osiągnięcia tych celów w opracowaniu przestawiono schemat (1) budowy wydajnego i bezpiecznego fundamentu działalności, wspierającego przyjęty model operacyjny[2] funkcjonowania GUS. Fundament działalności[3] powstaje dzięki starannemu wyselekcjonowaniu procesów i systemów, które należy zintegrować informacyjnie i funkcjonalnie oraz wprowadzając niezbędne standardy i mechanizmy referencyjności, interoperacyjności i bezpieczeństwa. Jak z tego wynika, oprócz dokumentu zawierającego inicjatywy strategiczne, GUS musi także mieć odpowiadający celom tych inicjatyw dobrze zdefiniowany model operacyjny swojego funkcjonowania. Zarówno model operacyjny, jak i fundament działalności powinny być tworzone i rozwijane zgodnie z ideami architektury korporacyjnej[4]. Strukturalne ujęcie tej kwestii przedstawiono na schemacie (1).

W najbardziej znanych metodach odwołujących się do dorobku architektury korporacyjnej zakłada się, że zaprojektowanie i wdrożenie fundamentu działalności organizacji (na poziomie teleinformatyki) wymaga opracowania: architektury procesów biznesowych, architektury danych, architektury aplikacji oraz architektury techniczno-systemowej. Jednocześnie z uwagi na znaczenie statystyki publicznej dla infrastruktury informacyjnej państwa, w tym przede wszystkim zasobów informacyjnych gromadzonych i przetwarzanych na po-

[1] Pojęcie bezpieczeństwa informacyjnego, na użytek tego artykułu, oznacza cechę systemu informacyjnego państwa polegającą na zapewnieniu obywatelom, przedsiębiorcom i administracji publicznej (szerzej państwu) informacji niezbędnych do skutecznego funkcjonowania zgodnego z obowiązującym prawem.

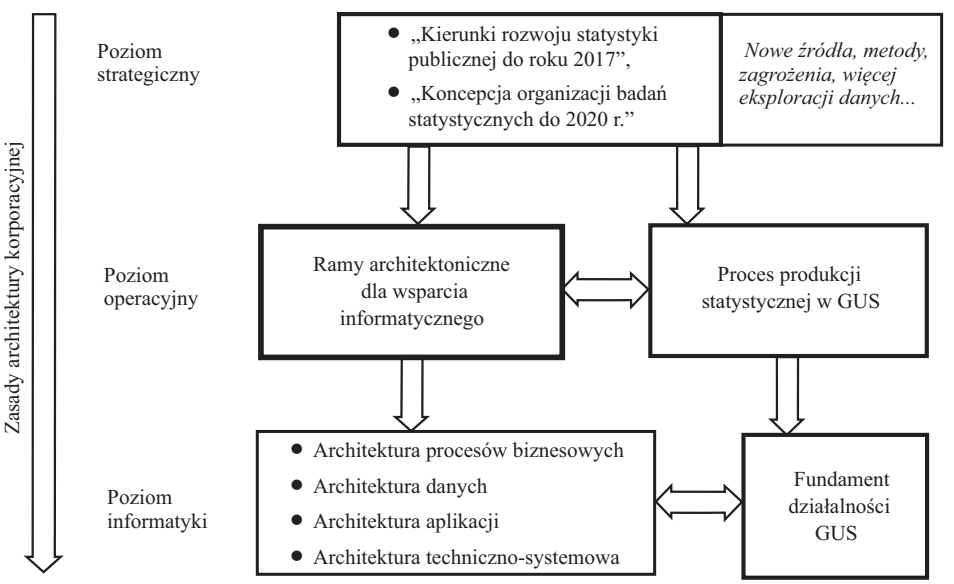
[2] Model operacyjny danej organizacji to opis, w jaki sposób ta organizacja chce osiągnąć zamierzone cele, w tym zwłaszcza oparte na wykorzystaniu narzędzi i metod informatyki.

[3] O organizacjach, które trwonią czas kierownictwa i inwestycje technologiczne (teleinformatyczne) na znaczną liczbę projektów mających uzasadnienie w skali lokalnej, lecz niestanowiących wsparcia dla nadrzędnych celów tych organizacji, mówi się, że nie mają wdrożonego fundamentu działalności.

[4] Architektura korporacyjna (ang. *Enterprise architecture*) to dziedzina wiedzy zajmująca się badaniem i opracowywaniem zasad i metod, których zastosowanie przyczynia się do skutecznej koordynacji i realizacji procesów zmierzających do osiągnięcia celów określonych w dokumentach strategicznych danej organizacji, zwłaszcza celów wymagających użycia technologii teleinformatycznych.

trzeby badań statystycznych, model operacyjny musi zapewnić wymagany poziom bezpieczeństwa danych statystycznych. W wymiarze praktycznym oznacza to potrzebę (szersze rozważania na ten temat wykraczają poza ramy artykułu) powrotu do koncepcji rozszerzenia tej listy o architekturę bezpieczeństwa informacyjnego, której własności powinny oddziaływać, zgodnie z zasadami projektowania mechanizmów bezpieczeństwa, na każdy etap rozwoju fundamentu działalności danej organizacji. Powinna ona uwzględniać niedocenianą, a nawet wręcz ignorowaną, najczęściej z braku świadomości specyfikę (odrębność)[5] metod ochrony statystycznych baz danych w stosunku do metod powszechnie implementowanych w transakcyjnych zastosowaniach technologii baz danych. Warto podkreślić, że dostępne na rynku uniwersalne systemy zarządzania bazami danych nie oferują mechanizmów, które w wymaganym stopniu wspierają mechanizmy ochrony statystycznych zastosowań baz danych. Brak jawnej, publicznej specyfikacji wspomnianej architektury bezpieczeństwa danych statystycznych skutkuje wdrażaniem rozwiązań zastępczych, nie gwarantujących wymaganego poziomu poufności.

SCHEMAT (1) POZIOMÓW OPISU STATYSTYKI PUBLICZNEJ



Ź r ó d ł o: opracowanie własne.

[5] W rzeczywistości chodzi o to, że stosowane w klasycznych zastosowaniach baz danych mechanizmy ochrony danych są niewystarczające i powinny być wzbogacone (uzupełnione) o warstwę mechanizmów uwzględniających specyfikę statystycznych baz danych.

Statystyka publiczna jest najważniejszym, niezależnym i profesjonalnym źródłem informacji o sytuacji społeczno-gospodarczej kraju. Z tego względu musi dysponować możliwościami gromadzenia i przetwarzania dla celów statystycznych ogromnych zasobów informacyjnych, co w konsekwencji generuje potrzebę zbudowania wydajnego i przede wszystkim bezpiecznego fundamentu wspierającego działalność GUS. Architektura bezpieczeństwa stanowiąca podstawę budowy skutecznych mechanizmów ochrony informacji statystycznych musi być rozpatrywana w pierwszym rzędzie na poziomie funkcjonalnym, czyli metod ochrony[6]. Tylko takie podejście pozwoli już w fazie analitycznej procesu projektowania (zgodnie z zasadą *security by design*) mechanizmów ochrony danych uniknąć zależności od wad dotychczasowych rozwiązań, a zwłaszcza od ograniczeń rozwiązań oferowanych w standardowych, komercyjnie dostępnych systemach zarządzania bazami danych. Biorąc to pod uwagę sformułowany wcześniej główny cel artykułu, warto zwrócić uwagę na zasadniczo odmienną problematykę bezpieczeństwa w statystycznych zastosowaniach technologii bazy danych w stosunku do szeroko znanej problematyki ochrony transakcyjnych baz danych, zwłaszcza na tle historii rozwoju mechanizmów ochrony danych. Chcąc uwypuklić potrzebę podjęcia i rozwijania przez GUS we współpracy ze środowiskami naukowymi prac nad metodami skutecznej ochrony danych statystycznych, świadomie zrezygnowano w artykule z formalnych zapisów (np. matematycznych) nie zawsze przejrzystych, skupiając się wyłącznie na problemowym ujęciu omawianych zagadnień. Dodatkowo Czytelnikom należy się wyjaśnienie, że autor artykułu nie wywodzi się ze środowiska statystyki publicznej i dlatego użyta w artykule terminologia może odbiegać od powszechnie stosowanej. Tym niemniej autor wyraża nadzieję, że wystarczająco jasno oddaje istotę poruszanych zagadnień.

PODSTAWOWE MODELE OCHRONY DANYCH W BAZACH DANYCH

Problem konieczności wyposażenia systemów komputerowych w rozbudowane mechanizmy ochrony danych dostrzeżono już w połowie lat sześćdziesiątych ub. wieku (Petersen i Turn, 1967), tj. jeszcze przed powstaniem technologii bazy danych. Mimo upływu kilkudziesięciu lat nadal aktualny jest zaproponowany przez Denning (w jej najważniejszej książce) podział mechanizmów ochrony danych w bazach danych na mechanizmy sterowania: dostępem, przepływem, szyfrowaniem i wnioskowaniem danych (Denning, 1982). Ograniczając zakres rozważań można zauważyć, że szyfrowanie (którym nie będziemy się zajmować), jako najstarsza technika ochrony poufności danych, zostało rozwinięte

[6] Mimo że już w pierwszej połowie XIX w. zaczęto traktować statystykę jako naukę (wyrazem tego było powołanie w 1834 r. w Anglii Królewskiego Towarzystwa Statystycznego oraz zwołanie w 1854 r. I Międzynarodowego Kongresu Statystycznego), osiągnięcia naukowe w dziedzinie metod ochrony statystycznych baz danych są mało znane i przede wszystkim nie znalazły zastosowania w mechanizmach ochronnych wdrożonych w komercyjnych systemach zarządzania bazami danych.

poza informatyką, głównie przez matematyków i specjalistów łączności. Z punktu widzenia systemów informatycznych, w tym systemów baz danych, stanowi ono przede wszystkim problem natury technicznej, polegający na poszukiwaniu sposobów wydajnej implementacji teoretycznie rozpoznanych metod szyfrowania. Warto również podkreślić, że technika ta zapobiega lub znacznie utrudnia ujawnienie treści informacji, natomiast nie zapobiega w ogóle zniekształceniu lub zniszczeniu danych i dlatego w bazach danych może być stosowana jedynie jako technika uzupełniająca.

Sterowanie dostępem do danych

Najważniejszy, powszechnie do dziś stosowany mechanizm sterowania dostępem do danych powstał na podstawie doświadczeń zebranych w trakcie realizacji i eksploatacji pierwszych inżynierskich rozwiązań mechanizmów ochrony danych w systemach operacyjnych. Istota funkcjonowania tego mechanizmu została opisana, uogólniona i przedstawiona przez Lampsona (1969) w postaci formalnego modelu znanego także pod nazwą modelu macierzy dostępu lub — w zależności od sposobu reprezentacji macierzy — modelu list zdolności (upoważnień). Podstawowymi elementami modelu są następujące zbiory: obiektów aktywnych zwanych podmiotami — S; obiektów pasywnych zwanych obiektami — O; przywilejów (praw) dostępu — T, które są trójkami (s, o, t), gdzie $s \in S$ i $o \in O$, a t jest zbiorem operacji, które podmiot s może użyć w stosunku do obiektu o. Zbiór reguł jest zwykle przedstawiony w postaci macierzy **A,** w której kolumny odpowiadają obiektom o_1, o_2, ...o_j, ...o_j ze zbioru O, wiersze podmiotom s_1, s_2, ...s_i, ...s_i ze zbioru S, a element macierzy dostępu $\mathbf{A}[s_i, o_j]$ jest przyznanym danemu podmiotowi przywilejem dostępu t_{ij}. W okresie od ogłoszenia modelu powstało wiele jego rozszerzeń dotyczących przede wszystkim rozszerzenia pojęcia przywileju dostępu poprzez uwzględnienie m.in.:

- identyfikatora osoby, która ma prawo (np. jako właściciel lub dysponent obiektów-danych) utworzyć, przyznawać, odbierać i zmieniać przywileje dostępu do jego obiektów;
- wskaźnika potwierdzającego fakt przeniesienia przez właściciela (dysponenta) prawa przyznawania, odbierania i zmieniania przywilejów dostępu do jego obiektów na inne osoby;
- dodatkowych kryteriów dostępu do obiektów (zwykle w postaci predykatu, np. uzależniającego udzielenie dostępu od wartości danych), ograniczenia okresu w którym dostęp jest możliwy czy wskazania miejsca z którego jest on możliwy itp.;
- warunków wywołania procedury realizującej określone czynności podczas upoważniania dostępu.

Niezależnie jednak od stopnia skomplikowania przywilejów dostępu reguła upoważniania dowolnego żądania takiego dostępu jest w tym modelu w swej istocie taka sama i polega na sprawdzeniu, czy dla danego podmiotu w macierzy znajduje się przywilej dostępu do określonego w żądaniu obiektu, dopuszczający realizację tego żądania.

Model sterowania przepływem

Model opracowany i opublikowany przez Denning (1976) jest najpełniejszym matematycznym modelem sterowania przepływem danych. Stanowi on formalne rozwiązanie problemu sterowania przepływem danych, który po raz pierwszy został opisany w modelu Bella i LaPaduli (1974)[7]. W modelu Denning uogólniono znane z modelu Bella i LaPaduli pojęcia kategorii i klasyfikacji, zastępując je pojęciem klasy ochrony. Opisowe reguły ochrony danych zastąpiono w nim jawnie zdefiniowaną relacją przepływu, określającą dopuszczalność przepływu danych między obiektami (zrezygnowano w nim z rozróżnienia pojęć podmiotów i obiektów znanych już z modelu Lampsona. W tym modelu używa się dla nich wspólnej nazwy „obiekt", bowiem są traktowane tak samo), stosownie do przydzielonych im klas ochrony. Oprócz zbioru klas ochrony i relacji przepływu, do podstawowych elementów modelu należą zbiory procesów i obiektów oraz operator „składania" klas ochrony, który służy do wyznaczania klasy ochrony wyniku będącego skutkiem wykonania dowolnego procesu lub jego części. Model *FM* sterowania przepływem został formalnie zapisany jako:

$$FM = [N, P, KO, OP, \rightarrow]$$

gdzie:

N — zbiór obiektów mogących przechowywać lub otrzymywać dane,
P — zbiór procesów przetwarzania powodujących przepływy danych,
KO — zbiór klas ochrony,
OP — operator „składania" klas ochrony,
→ — relacja przepływu zdefiniowana na parach klas ochrony.

Dla tak sformułowanego modelu autorzy Denning (1976) i Szafrański (1987) wykazali na podstawie analizy semantyki przetwarzania danych, że sensowne jest przyjęcie założenia, iż elementy modelu, tj. zbiór klas ochrony oraz relacja przepływu generują algebraiczną strukturę kraty. Ta krata, nazwana w modelu

[7] Model ten w swej podstawowej postaci zawiera zbiory: elementów aktywnych (podmiotów), obiektów pasywnych (obiektów), kategorii i klasyfikacji oraz pojęcie bieżącego i bezpiecznego stanu systemu. Podmioty i obiekty mają takie samo znaczenie, jak w modelu Lamspona. Klasyfikacje najczęściej odnoszą się do przydzielonych podmiotów i obiektów poziomu tajności. Natomiast kategorie i ich przydział do podmiotów i obiektów służą do odzwierciedlenia określonego przez projektanta ich podziału na rozłączne klasy (np. według przynależności do określonych sfer działalności instytucji). Poziomy tajności i podzbiory zbioru kategorii tworzą łącznie pary nazwane poziomami ochrony. Bardzo istotną cechą modelu jest to, że przyjęto w nim częściowe uporządkowanie zbioru poziomów ochrony. Przyjęto także, iż przy pewnych praktycznie sensownych założeniach częściowo uporządkowany zbiór generuje strukturę kraty, w której istnieją operatory kresów górnych i dolnych. W związku z tym model ten jest często nazywany wielopoziomowym kratowym modelem ochrony danych. W modelu tym zakłada się, że bieżące stany systemu ulegają zmianie pod wpływem tzw. żądań, które mogą być żądaniami dostępu do obiektów zmieniającymi przyznany poziom ochrony itd. Uwzględniając to, określono warunki bezpiecznej zmiany stanu systemu, tzn. nienaruszającej zasad ochrony.

kratą przepływu, jest podstawowym elementem kratowego modelu sterowania przepływem danych, w którym operatory kresów górnego i dolnego o ściśle matematycznie zdefiniowanych własnościach mogą być jednoznacznie zaimplementowane w praktycznych mechanizmach ochrony danych.

Najważniejsze jednak z punktu widzenia tego artykułu jest podkreślenie, że:

- oba omówione wyżej modele koncentrują się na kontroli dostępu do wartości atrybutów indywidualnych (egzemplarzowych) obiektów przechowywanych w bazie danych. W szczególności oznacza to dostęp do danych dotyczących poszczególnych osób;
- polityka ochrony danych w tych modelach, najogólniej pisząc, polega na zabronieniu realizacji żądań niezgodnych ze stanem macierzy dostępu lub powodujących przepływy danych niezgodne z określoną w modelu relacją przepływu.

OCHRONA DANYCH W STATYSTYCZNYCH BAZACH DANYCH

Na konferencji naukowej, która odbyła się w 2015 r. w GUS, sformułowano[8] następujące pytania nawiązujące do treści niniejszego artykułu:

- czy dla opisu problemów ochrony statystycznych baz danych konieczne jest rozszerzenie bazy terminologicznej,
- czy dla ochrony statystycznych baz danych niezbędne są nowe metody ochrony danych, odmienne od dotychczas stosowanych w komercyjnie dostępnych systemach zarządzania bazami danych.

Uczestnicy konferencji zarówno w dyskusji plenarnej, jak i w licznych dyskusjach kuluarowych potwierdzili zasadność tych pytań, jednocześnie wskazując na dodatkowe argumenty wynikające przede wszystkim z potrzeby uwzględnienia w badaniach statystycznych (o ile prawo na to zezwoli) nowych źródeł danych (zwłaszcza nieustrukturyzowanych danych internetowych) oraz metod Big Data[9] (np. nieomawianych w tym artykule zagadnień scalania danych z różnych źródeł, rozwiązywania konfliktów wiarygodności źródeł danych, próbkowania wiarygodności danych w szybkozmiennym środowisku wymiany danych itp.).

Należy zauważyć, że wcześniej przedstawiono podstawowe modele ochrony danych w bazie danych niezależnie od ich dalszego rozwoju (skomplikowania), które spełniają następujące **reguły sterowania:**

- **dostępem** — upoważnienie dowolnego żądania (operacji) dostępu do danych polega na sprawdzeniu w macierzy, czy istnieje przywilej (prawo) dostępu dopuszczający jego realizację;
- **przepływem** — upoważnienie przepływu danych spowodowane dowolnym żądaniem (operacją) dostępu do danych polegające na sprawdzeniu, czy w wy-

[8] Konferencja *System Informacyjny Statystyki Publicznej wobec wyzwań prawnych, naukowych i technologicznych w obszarze ochrony danych statystycznych*, GUS, 2015 r., Warszawa.

[9] Od niedawna komitet językowy do spraw nazewnictwa informatycznego zaleca używać polskiego odpowiednika angielskiego Big Data zamiast terminu *gigadane*.

niku jego realizacji nie nastąpi przepływ danych niezgodny z określoną relacją przepływu, czyli od obiektu o wyższej klauzuli poufności do obiektu o niższej klauzuli poufności.

Tak określone reguły ochrony danych w swej istocie są skupione na kontroli dostępu obiektów aktywnych do indywidualnych egzemplarzowych zapisów w bazie danych. Zgodnie z tymi regułami zbiór wszystkich indywidualnych danych jest dzielony na dwa rozłączne podzbiory, tj. danych dostępnych zgodnie z przyznanymi przywilejami (prawami) dostępu i niedostępnych dla danego użytkownika (programu).

Jeśli więc porówna się powyższe reguły z obowiązującą regułą w sterowaniu wnioskowaniem (Denning, 1992) w brzmieniu: *reguła sterowania wnioskowaniem — dostęp do danych w statystycznej bazie danych jest ograniczony do udostępniania danych wyłącznie w postaci statystyki*[10] *do tylko takich, które nie doprowadzą bezpośrednio lub pośrednio do ujawnienia danych dotyczących pojedynczych, zidentyfikowanych obiektów informacyjnych bazy danych (np. konkretnej osoby, transakcji czy podmiotu, ...)*, to z łatwością można zauważyć fundamentalne różnice między regułami ochrony w klasycznych (transakcyjnych, ewidencyjnych itp.) i statystycznych zastosowaniach baz danych[11]. Ta różnica musi skutkować potrzebą stosowania w części odmiennej terminologii, ale przede wszystkim potrzebą prowadzenia prac naukowych i wdrożeniowych, których celem jest opracowanie metod skutecznie zapobiegających ujawnianiu danych indywidualnych z zasobów utworzonych na potrzeby badań statystycznych. Jest to zagadnienie bardzo trudne, ponieważ trzeba się zgodzić z twierdzeniem, że każda statystyka, nawet zagregowana dla całej populacji, zawiera ślad informacji pierwotnych, indywidualnych.

W celu przedstawienia, choćby sygnalnie, głównych kierunków poszukiwań efektywnych metod ochrony danych w statystycznych bazach danych konieczne jest wprowadzenie kilku wybranych pojęć oddających specyfikę statystycznych baz danych. Należą do nich m.in. pojęcia:

- **statystyczna baza danych** — utworzona na potrzeby badań statystycznych, której celem jest gromadzenie, przetwarzanie i udostępnianie danych wyłącznie w postaci dopuszczonego zbioru statystyk;
- **uproszczony model statystycznej bazy danych** — zgodnie z tym modelem baza danych jest zbiorem N wierszy (rekordów), z których każdy składa się z wartości M atrybutów, przy czym każdy z atrybutów przyjmuje wartość z określonej dla niego dziedziny wartości (np. z dziedziny „wiek”);
- **stan informacyjny statystycznej bazy danych** — odzwierciedla wartości przechowywanych danych oraz wiedzy dodatkowej obejmującej wiedzę zewnętrzną i roboczą;

[10] Na potrzeby tego artykułu przyjęto za *Słownikiem Języka Polskiego PWN*, że statystyka jest liczbą lub zbiorem informacji liczbowych dotyczących jakichś zjawisk. W tym ujęciu znaczy to tyle, co wynikowe informacje statystyczne lub produkty statystyczne.

[11] W artykule nie rozważano problemów i metody anonimizacji danych.

- **wiedza zewnętrzna** — to wiedza dotycząca wartości atrybutów danych przechowywanych w statystycznej bazie danych, jednak nie jest ona uzyskana (bo nieudostępniana przez system statystycznej bazy danych) z odpowiedzi udzielanych przez statystyczną bazę danych;
- **wiedza robocza** — dotyczy listy obiektów i atrybutów, których wartości są przechowywane w statystycznej bazie danych (nazwy obiektów i atrybutów są jawne, natomiast ich wartości są poufne);
- **formuła charakterystyczna** — dowolne dla danego obiektu wyrażenie logiczne zbudowane z wartości atrybutów połączonych operatorami *or* (+), *and* (*), *not* (~); jeśli w statystycznej bazie danych znajdują się wartości atrybutów „zawód" i „płeć", to formuła może przyjąć np. postać: (*Zawód=Lekarz*)+(*Zawód= =Weterynarz*)*(*Płeć=Mężczyzna*);
- **zbiór odpowiedzi** — to zbiór wierszy (rekordów), którego wartości atrybutów spełniają określoną formułę charakterystyczną;
- **statystyka wrażliwa** — to statystyka, która ujawnia poufne informacje o pewnych obiektach; oczywiście dla zbioru odpowiedzi o liczności 1 statystyka jest zawsze wrażliwa. Jak z tego wynika, statystyka wrażliwa nie odnosi się do katalogu kategorii danych podlegających ochronie prawnej, lecz do wszystkich przypadków, w których jej upublicznienie prowadzi do ujawnienia informacji dotyczących pojedynczego obiektu obserwacji;
- **ujawnienie statystyczne** — sytuacja, w której użytkownik na podstawie zbioru dostępnej statystyki i posiadanej wiedzy dodatkowej może poprzez wnioskowanie uzyskać informacje o statystyce zastrzeżonej (wrażliwej). Takie rozumienie ujawnienia statystycznego wiąże się z zagadnieniem znanym w statystyce publicznej pod nazwą tajemnicy zwrotnej;
- **„kompromitacja"**[12] **statystycznej bazy danych** — jest to sytuacja, w której użytkownik na podstawie zbioru dostępnej statystyki i posiadanej wiedzy dodatkowej jest w stanie na podstawie wnioskowania uzyskać przechowywaną w statystycznej bazie danych informację dotyczącą wartości atrybutów pojedynczego obiektu.

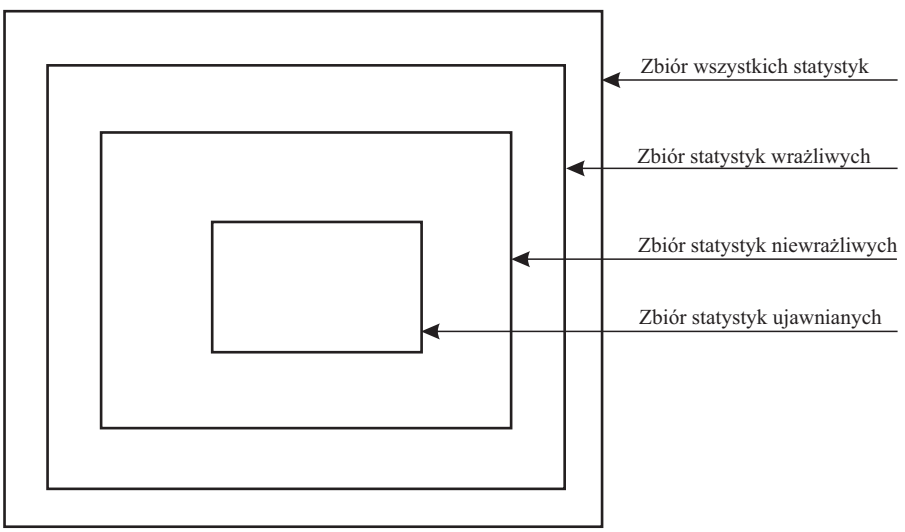
Biorąc pod uwagę przedstawione pojęcia oraz zilustrowane na schemacie (2) zależności miedzy różnymi rodzajami statystyk można wyróżnić m.in. następujące stany, w których **statystyczna baza danych jest:**

- **bezpieczna,** jeśli zbiór statystyk ujawnianych jest podzbiorem zbioru statystyki niewrażliwej, czyli wtedy gdy statystyczna baza danych nie udostępnia żadnej statystyki wrażliwej;
- **zupełna** (wyczerpująca, dokładna), jeśli zbiory statystyki udostępnianej i niewrażliwej są równe (takie same).

[12] Autor zdaje sobie sprawę, że terminy „kompromitacja", „skompromitowana" nie są stosowane w statystyce publicznej. W artykule użyto ich jako polskich odpowiedników angielskich *compromising* i *compromised*. Terminy te mają zbliżone znaczenie do pojęcia „ujawnienie danych jednostkowych w wyniku naruszenia tajemnicy statystycznej".

Bardzo istotnym problemem w statystycznych bazach danych są wzajemne zależności między udostępnianą statystyką. Może się bowiem okazać, że udostępnienie pewnej statystyki może doprowadzić do „uwrażliwienia" wcześniej już udostępnionej, a traktowanej jako niewrażliwą, statystykę. Jednak nadmiernie rygorystyczna polityka udostępniania statystyki może doprowadzić do utraty cechy zupełności statystycznej bazy danych poprzez profilaktyczne, „nadgorliwe" (na wszelki wypadek) zabronienie udostępniania statystyk, które w swej istocie nie stwarzają zagrożeń udostępnienia statystyk wrażliwych. Złożoność przedstawianego tu problemu wynika przede wszystkim z trudności jednoznacznego stwierdzenia zarówno faktu, jak i rozległości wystąpienia tego rodzaju powiązań między statystykami, bowiem zawsze powstaje jakieś ryzyko (i nie da się go całkowicie wyeliminować).

SCHEMAT (2) ZALEŻNOŚCI MIĘDZY RODZAJAMI STATYSTYK



Ź r ó d ł o: na podstawie pracy Denning (1982).

KLASYFIKACJA METOD ATAKU I OCHRONY STATYSTYCZNEJ BAZY DANYCH

Metody ataku na statystyczne bazy danych

Właściwe i dogłębne zrozumienie reguły ochrony statystycznych baz danych jest bardzo pomocne przy rozważaniu potencjalnych zagrożeń dla bezpieczeństwa danych statystycznych. Warto jeszcze raz zwrócić uwagę na cel ataków

(poznanie wartości atrybutów konkretnego indywidualnego obiektu) oraz na szczególne znaczenie tzw. zewnętrznej wiedzy dodatkowej, która zwiększa szanse wydedukowania informacji wrażliwej na podstawie legalnie uzyskanych statystyk. Rozpatrując zagrożenia dotyczące bezpieczeństwa danych statystycznych należy mieć na uwadze nie tylko to, że bezpośrednie uzyskanie wrażliwych danych ze statystycznej bazy danych oznacza jej „kompromitację", ale również to, iż do uzyskania takich danych można doprowadzić wnioskując na podstawie legalnie udostępnianej statystyki oraz posiadanej dodatkowej wiedzy zewnętrznej. Opierając się w znacznej mierze na wspomnianej już publikacji Denning oraz dorobku konferencji (Domingo-Ferrer i Magkos, 2010) przedstawiono — w celu uzasadnienia tezy o zasadniczej odmienności metod ochrony statystycznych baz danych — bardzo skrótowo samą istotę przykładowych dwóch metod ataków i ochrony charakterystycznych dla statystycznych zastosowań baz danych. Pozostałe znane z literatury metody zostały wymienione tylko z nazwy, ponieważ już same nazwy świadczą dobitnie o ich odrębności w stosunku do metod ochrony klasycznych. Ponadto pragnę zasygnalizować, że pełniejszemu opisowi oraz analizie tych metod będzie poświęcony oddzielny artykuł.

Metoda oparta na powiększaniu liczebności zbioru odpowiedzi

Metodę tę stosuje się najczęściej w sytuacji, gdy baza danych jest chroniona mechanizmami wykorzystującymi opisaną wcześniej technikę sterowania liczebnością zbioru odpowiedzi spełniających zadaną formułę charakterystyczną. Znajduje ona zastosowanie przede wszystkim w przypadku małego zbioru odpowiedzi oraz dopuszczenia możliwości uzupełnienia zawartości bazy danych o liczbę wierszy (rekordów) spełniających charakterystyczną formułę niezbędną do osiągnięcia liczebności zbioru odpowiedzi, zezwalającej na udzielanie odpowiedzi. W celu zautomatyzowania procesu uzyskiwania wartości atrybutów konkretnych obiektów (np. osób) opracowano i udowodniono (Schlorer, 1976) przydatność i formalną poprawność działania tzw. szperaczy, wśród których wyróżniono szperacze: indywidualne, ogólne, podwójne i łączne. Znalezienie i zastosowanie szperacza pozwala wrażliwą (niedozwoloną) statystykę obliczyć po zadaniu zaledwie kilku legalnych pytań. Co więcej, w opracowaniu (Dening i Schlorer, 1980) opublikowano algorytm, który dla określonej charakterystyki bazy danych podaje jednoznaczną procedurę automatycznego znajdowania stosownych szperaczy.

Metoda oparta na analizie bardzo małych zbiorów odpowiedzi

W wielu pracach (Hoffman i Miller, 1970) poświęconych bezpieczeństwu danych statystycznych wykazano, że istnieją łatwe sposoby „skompromitowania" (czasem używa się pojęcia „przenikania") statystycznej bazy danych w sytuacji dopuszczenia do ujawniania statystyk obliczonych na podstawie zbiorów odpo-

wiedzi o małej liczbie wierszy w stosunku do ich ogólnej liczby w bazie danych. Najprostszym przykładem takiej analizy jest sytuacja, w której intruz wie na pewno na podstawie dodatkowej wiedzy zewnętrznej, że informacje dotyczące osoby *X* znajdują się w bazie danych i odpowiadają danej formule charakterystycznej *Y*. Jeśli więc na pytanie o liczebność zbioru odpowiedzi dla tej formuły uzyska odpowiedź, że równa się ona 1, to ma pewność, że doprowadził do jej identyfikacji i może dalej prowadzić penetrację w celu uzyskania informacji, czy *X* ma dodatkową cechę *Z*. Może to zrobić zadając pytanie o wielkość zbioru odpowiedzi dla statystyki o liczebności (*Y***Z*). Oczywiście odpowiedź 1 oznacza, że *X* ma tę cechę, a 0, że jej nie ma. Obie uzyskane odpowiedzi świadczą o przeniknięciu do bazy danych. Biorąc to pod uwagę statystyki obliczane na podstawie zbyt małych zbiorów odpowiedzi powinny być zakwalifikowane jako wrażliwe, jednak nie zawsze w praktyce tak się postępuje.

Inne metody ataku na statystyczne bazy danych

Do tej grupy należą metody oparte na:

- tworzeniu i rozwiązywaniu układu równań matematycznych,
- pytaniach o charakterystyki statystyczne (ataki medianą, estymatory statystyk wrażliwych),
- wykorzystaniu cech dynamicznych baz danych,
- analizie statystyk mających pokrywające się zbiory odpowiedzi (jest ona szczególnie groźna, gdy stopień pokrywania jest duży),
- analizie statystyk specyfikowanych przez klucze.

Metody ochrony statystycznych baz danych

Metoda oparta na sterowaniu liczebnością zbioru odpowiedzi

Jest to metoda niezwykle prosta, bardzo łatwa w implementacji i nieznacznie obciążająca czas obliczania statystyk oraz niedopuszczająca do ich udostępniania, jeśli zostały obliczone na mniejszym zbiorze odpowiedzi niż to wynika z ustalonej dla tej statystyki polityki dostępu. Użyteczność tej metody zależy w znacznym stopniu od sposobu liczenia wartości minimalnej liczebności zbioru odpowiedzi do obliczenia danej statystyki. Nieuzasadnione nadmierne podwyższanie granicznej liczebności prowadzi co prawda do zwiększenia bezpieczeństwa danych statystycznych, ale może jednocześnie istotnie zmniejszyć użyteczność takiej bazy danych. Analiza ryzyka w tym przypadku musi uwzględniać zarówno kwestie wynikające z profilu wykorzystywania statystyki, jak i prawdopodobieństwa posiadania przez potencjalnych intruzów użytecznej dla nich w procesie penetracji wiedzy dodatkowej oraz prawdopodobieństwa stosowania określonych metod „kompromitacji" bazy danych.

Metoda oparta na przekształcaniu (zniekształcaniu) odpowiedzi lub wartości atrybutów

Metoda ta należy do budzących największe zainteresowanie z uwagi na swą prostotę i jednocześnie wysoką skuteczność, zwłaszcza gdy intruzi dysponują dodatkową wiedzą zewnętrzną. W największym stopniu przeciwdziała ona atakom z wykorzystaniem szperaczy lub układów równań liniowych. Zniekształcanie polega na wprowadzeniu tzw. szumów do statystyki poprzez stosowanie różnych metod zaokrąglania lub zaszczepiania szumem wartości atrybutów w taki sposób, by mimo losowego „zaszumienia" pierwotnych wartości udostępniana statystyka była wystarczająco dokładna. Wprowadzanie szumu do wartości atrybutów przechowywanych w statystycznej bazie danych może mieć charakter stały poprzez trwałe zmodyfikowanie ich wartości lub występować w momencie wyliczania na ich podstawie określonej statystyki. Sterowanie tym procesem należy do zadań mechanizmu ochrony zaimplementowanego w statystycznej bazie danych.

Lista innych metod ochrony statystycznej bazy danych

Do tej grupy należą metody oparte na:
- podziale bazy danych,
- losowaniu zbioru odpowiedzi,
- zamianie wartości atrybutów,
- analizie stopnia pokrywania się zbiorów odpowiedzi,
- ograniczaniu dopuszczalnych statystyk (sterowania rzędem statystyki) lub zabranianiu komórek.

Wnioski

Odnosząc się do tytułu artykułu należy stwierdzić, że z uwagi na rosnące znaczenie analiz statystycznych oraz eksploracyjnych cyfrowych zasobów informacyjnych należy bezwzględnie zastosować w Polsce powszechnie uznane, teoretyczne i praktyczne podstawy budowy bezpiecznych mechanizmów ochrony statystycznych baz danych. Jest to najważniejszy wniosek i jednocześnie najważniejsze przesłanie artykułu. Do obsługi (lub wsparcia) badań statystycznych konieczne jest stosowanie uniwersalnych systemów zarządzania bazami danych wyposażonych w standardowe, nieuwzględniające specyfiki statystycznych zastosowań bazy danych. Co więcej, o ile można mówić o wzmożonych pracach skupiających wysiłki na identyfikacji i badaniu zagrożeń w szeroko pojętym środowisku internetowym, to prace dotyczące ochrony danych statystycznych w znacznej mierze w Polsce mają charakter fragmentaryczny odległy od możliwości komercyjnego wdrożenia uzyskanych wyników.

Sformułowane tu uwagi dotyczą głównie polskiej statystyki publicznej. Nie oznacza to jednak, że w trakcie prowadzonych badań statystycznych, zwłaszcza tych wykorzystujących wsparcie informatyczne, bagatelizowany jest problem bezpieczeństwa danych statystycznych. Wskazują one natomiast na potrzebę inicjowania i prowadzenia prac interdyscyplinarnych podejmowanych wspólnie przez GUS i uczelniane zespoły badawcze. Ich celem powinno być zbadanie faktycznego ryzyka w zakresie bezpieczeństwa danych statystycznych w przypadku stosowania dostępnych na rynku uniwersalnych systemów zarządzania bazami danych. W przypadku potwierdzenia tezy o niedoskonałości standardowych mechanizmów ochrony należy wskazać rozwiązania, które powinny wzbogacić istniejące mechanizmy ochrony danych i w rezultacie sprowadzić wspomniane ryzyko do akceptowalnego poziomu.

Na zakończenie konieczne jest przedstawienie wyjaśnienia dotyczącego powoływania się w artykule na stosunkowo odległą w czasie literaturę dziedzinową. Jest to świadomy zabieg autora, który poprzez wskazanie pierwotnych źródeł (nawet z lat osiemdziesiątych ub. wieku) odnoszących się do przedstawionych w artykule problemów i metod chciał w ten sposób podkreślić, że nowsze publikowane wyniki tylko fragmentarycznie odnoszą się do przedstawionych problemów i nie tworzą spójnych podstaw teoretycznych i metodycznych w zakresie rozwoju i implementacji kompleksowych metod (mechanizmów) ochrony statystycznych baz danych. Budowa zwartej koncepcji (teorii) ochrony statystycznych baz danych powinna opierać się, zdaniem autora, przede wszystkim na pierwotnych źródłach literatury.

dr hab. inż. Bolesław Szafrański — profesor *Wojskowej Akademii Technicznej*

LITERATURA

Bell, D., LaPadula, L. (1974). *SecureComputer Systems: A Mathematical Foudations and Model*. Bedford. Stany Zjednoczone: MITRE Corporation.

Denning, D. (1976). A Lattice Model of Secure Information Flow. *CACM, vol. 19, no. 5*. Stany Zjednoczone.

Denning, D. (1982). *Cryptography and Data Security*. Stany Zjednoczone: Addison-Wesley. Wydanie polskie w 1992 r. Warszawa: Wydawnictwa Naukowo-Techniczne. Tłumaczenia — zespół pod kierunkiem B. Szafrańskiego.

Denning, D. (1992). *Kryptografia i ochrona danych*. Warszawa: Wydawnictwo Naukowo-Techniczne.

Denning, D., Schlorer, J., (1980). A fast procedure for finding a tracker in a statistical database. *Journal ACM Transactions on Database Systems, vol. 5, no. 1*, s. 88—102.

Domingo-Ferrer, J., Magkos, E. (eds.) (2010). *Privacy in Statistical Databases*. Corfu, Grecja.

Hoffman, L.J., Miller, W.F. (1970). Getting a personal dossier from a statistical data bank. *Datamation, vol. 16, no. 5*. Stany Zjednoczone.

Lampson, B.W. (1969). *Dynamic Protection Structures*. Stany Zjednoczone: Proceedings of AFIPS.

Petersen, H.E., Turn, R. (1967). *System Implementation of Information Privacy*. Stany Zjednoczone: Proceedings of AFIPS.

Schlorer, J. (1976). Identification and retrieval of personal records from a statistical data bank. Methods Information Mrd. *Journal ACM Transactions on Database Systems, vol. 5, no. 1*. Stany Zjednoczone.

Szafrański, B. (1987). *Modelowanie procesów ochrony bazy danych ze szczególnym uwzględnieniem ich integracji*. Warszawa: Oficyna WAT.

Summary. *The main aim of this article was to highlight to the importance of research and experimental development studies concerning methods for the protection of statistical databases, when the universal database management systems do not provide the mechanisms supporting the required security level of statistical bases. The study presents the basic models of controlling access and data flow, and it proves their limited relevance to statistical databases. Moreover, the specific nature of security issues as well as methods of attacking and protecting such databases were discussed.*

In conclusion, it was stated that Poland does not apply universally recognized, theoretical basics for the development of secure protection mechanisms of statistical databases. The work on data protection was fragmentary, distant from the possibilities of commercial implementation of the results. Therefore, there is a need for interdisciplinary work of the CSO and academic research teams.

Keywords: database, data protection, controlling: access, flow and inference, statistical database.