Dariusz Dymek

Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej
Akademia Górniczo-Hutnicza w Krakowie

Katedra Systemów Obliczeniowych
Akademia Ekonomiczna w Krakowie

Wojciech Komnata, Leszek Kotulski

Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej
Akademia Górniczo-Hutnicza w Krakowie

Federacyjna hurtownia danych w dostępie do informacji poufnej[1]

1. Wstęp

Organy ochrony prawa na potrzeby wykonywania swoich ustawowych obowiązków tworzą bazy danych gromadzące m.in. informacje o: osobach, organizacjach, podmiotach gospodarczych czy zdarzeniach. Informacje te mają charakter poufny, a ich zakres i sposób użycia jest regulowany przepisami prawa. Jednym z dopuszczalnych sposobów wykorzystania tych danych jest wzajemne udostępnianie danych przez poszczególne służby zaliczane do organów ochrony prawa w zakresie niezbędnym do realizacji zadań, jakie przed nimi są postawione.

Poufny charakter danych sprawia, że wszelkie możliwości ich gromadzenia i udostępnienia podlegają regulacjom prawnym, a sam proces gromadzenia

[1] Praca jest częścią projektu badawczego prowadzonego przez Katedrę Informatyki Stosowanej Wydziału Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej Akademii Górniczo-Hutniczej w Krakowie „Nowoczesne technologie dla/w procesie karnym i ich wykorzystanie – aspekty techniczne, kryminalistyczne, kryminologiczne i prawne", nr umowy: 0021/R/ID2/2011/01, Nr AGH: 17.17.120.161.

i udostępniania danych jest formalnie określony i podlega nadzorowi. Niemniej jednak co jakiś czas w mediach mówi się o wątpliwościach dotyczących tego, czy nadzór ten jest realizowany właściwie. Jedną z ich przyczyn jest fakt, że ze względu na poufny charakter tych danych bezpośredni nadzór nad ich wykorzystaniem sprawują te same organy ochrony prawa, które tymi danymi się posługują w swojej pracy, a nadzór zewnętrzny jest sprawowany *post factum* i często obejmuje jedynie informacje o samym zgromadzeniu lub wymianie danych, bez możliwości weryfikacji prawdziwości przesłanek podjęcia tych działań.

Jednym z powodów takiej sytuacji jest ograniczony zakres wykorzystywania technologii informatycznych w procesie udostępniania danych. Chociaż służby dysponują własnymi wewnętrznymi rozwiązaniami opartymi na technologii informatycznej, często bardzo zaawansowanymi, to w procesie wymiany danych, a w szczególności danych poufnych, pomiędzy organami ochrony prawa dominują procedury wykorzystujące tradycyjną formę papierową. Budowa systemu informatycznego pozwalającego w sposób efektywny udostępniać sobie wzajemnie dane przy jednoczesnym zachowaniu wymaganego poziomu bezpieczeństwa i możliwości kontroli jest wyjątkowo utrudniona. Wynika to z wielu czynników, wśród których na pierwszym miejscu niewątpliwie należy wymienić uwarunkowania prawne i technologiczne oraz ich wzajemną interakcję. Jako przykład wzajemnej interakcji tych uwarunkowań można wskazać wielość baz danych o różnorakiej strukturze zbudowanych przy wykorzystaniu różnych narzędzi. Problem ten może być technologicznie skutecznie rozwiązany dzięki mechanizmom zarządzania heterogenicznymi bazami danych[2]. Niestety takie rozwiązanie wymaga znajomości (a co za tym idzie częściowego upublicznienia) struktur danych wszystkich baz danych wchodzących w skład systemu, a to z prawnego punktu widzenia jest w wielu wypadkach niemożliwe. Ponadto takie otwarcie baz danych mogłoby stanowić poważne zagrożenie dla bezpieczeństwa danych, które z założenia mają charakter poufny. Z tego też powodu takie rozwiązanie jest nie do przyjęcia.

Już na tym przykładzie można ukazać podstawowe problemy dotyczące stworzenia systemu informatycznego wspierającego wymianę danych poufnych. System taki powinien dostarczyć mechanizmów wzajemnego udostępniania danych, ale jedynie w zakresie wynikającym z przepisów prawa, przy zachowaniu wszelkich wymogów bezpieczeństwa, m.in. bez ujawniania szczegółów budowy

[2] R. Wrembel, *Usługi heterogeniczne – techniki integracji rozproszonych baz danych różnych producentów*, Materiały konferencyjne XI Konferencji PLOUG „Systemy informatyczne: projektowanie, implementowanie, eksploatowanie", 18–21 października, Zakopane–Kościelisko 2005, s. 327–352; L.M. Haas, E.T. Lin, M.A. Roth, *Data integration through database federation*, „IBM Systems Journal" 2002, vol. 41, no. 4, s. 578–596.

baz danych, oraz zapewnić możliwość nadzoru nad procesem wymiany danych zarówno na poziomie wewnętrznym, jak i zewnętrznym.

Niniejsza praca prezentuje koncepcję wykorzystania mechanizmów opartych na federacyjnej hurtowni danych do stworzenia systemu wymiany poufnych danych spełniającego powyższe wymagania. Paragraf 2 stanowi ogólną charakterystykę hurtowni danych ze szczególnym uwzględnieniem tych mechanizmów, które będą później odgrywać istotną rolę w prezentowanej koncepcji. W kolejnej części omówiono aspekty prawne wymiany danych oraz wskazano uwarunkowania wyboru architektury prezentowanego rozwiązania. W paragrafie 4 przedstawiono koncepcję i architekturę systemu wymiany danych poufnych. W następnej części przedstawioną koncepcję systemu rozszerzono o elementy wspierające działania kontrolne nad procesem wymiany danych oparte na mechanizmach federacyjnej hurtowni danych. Na zakończenie podsumowano uzyskane rezultaty oraz wskazano kierunki dalszych badań.

2. Ogólna koncepcja federacyjnej hurtowni danych

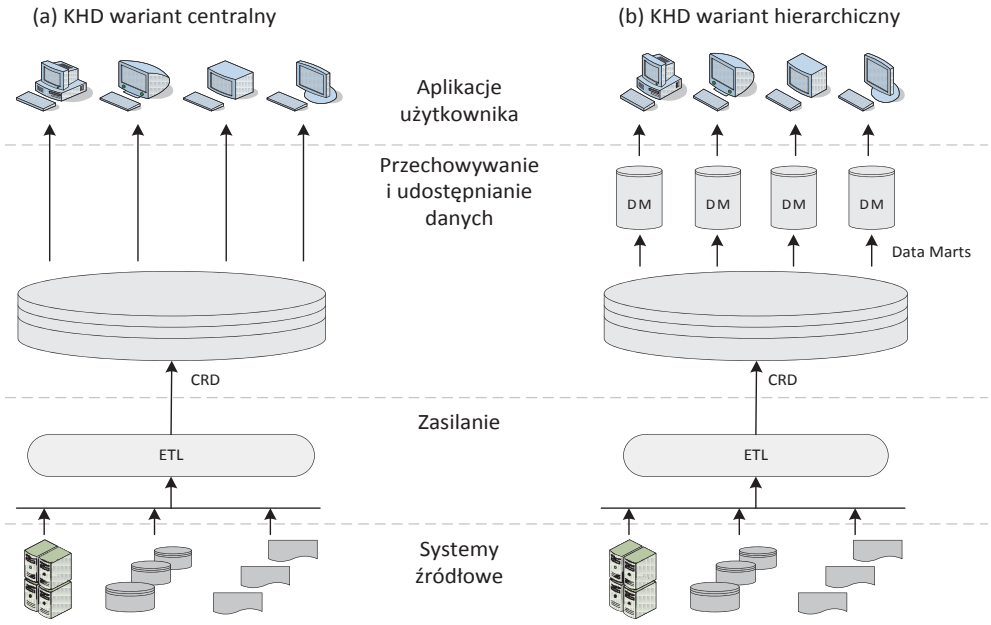
Koncepcja hurtowni danych, definiowanej jako „tematyczna baza danych, która trwale przechowuje zintegrowane dane opisane wymiarem czasu”[3], powstała w latach 70. XX w. w odpowiedzi na zapotrzebowanie głównie dużych korporacji w zakresie przechowywania i przetwarzania danych historycznych na potrzeby analiz biznesowych. W początkowym okresie wykorzystania hurtowni danych dominowała architektura systemu oparta na centralnym repozytorium danych (CRD), bazie danych zawierającej zintegrowane dane pochodzące z wielu źródeł. Taka architektura hurtowni danych, określana często jako korporacyjna hurtownia danych (KHD), zakłada fizyczne kopiowanie danych z systemów źródłowych. W procesie zasilania (ETL[4]) dane są integrowane i transformowane do modelu danych hurtowni (model danych CRD), ładowane do CRD, a następnie dane te są udostępniane aplikacjom użytkowników (rysunek 1, część a).

Rozwój wykorzystania hurtowni danych, a w szczególności rosnące wymagania w zakresie wydajności, doprowadził do rozbudowy architektury KHD o warstwę hurtowni tematycznych (ang. *Data Marts*). Taka architektura, nazywana

[3] W.H. Inmon, *Building the Data Warehouse*, John Wiley & Sons, New York 2005.

[4] *Extraction, Transforming and Loading* (ETL) – grupa procesów zasilających hurtownie danych odpowiedzialną za pozyskanie danych z systemów źródłowych, ich integrację i transformację do modelu danych hurtowni, weryfikację poprawności („czyszczenie” danych) oraz załadowanie tak przygotowanych danych do CRD.

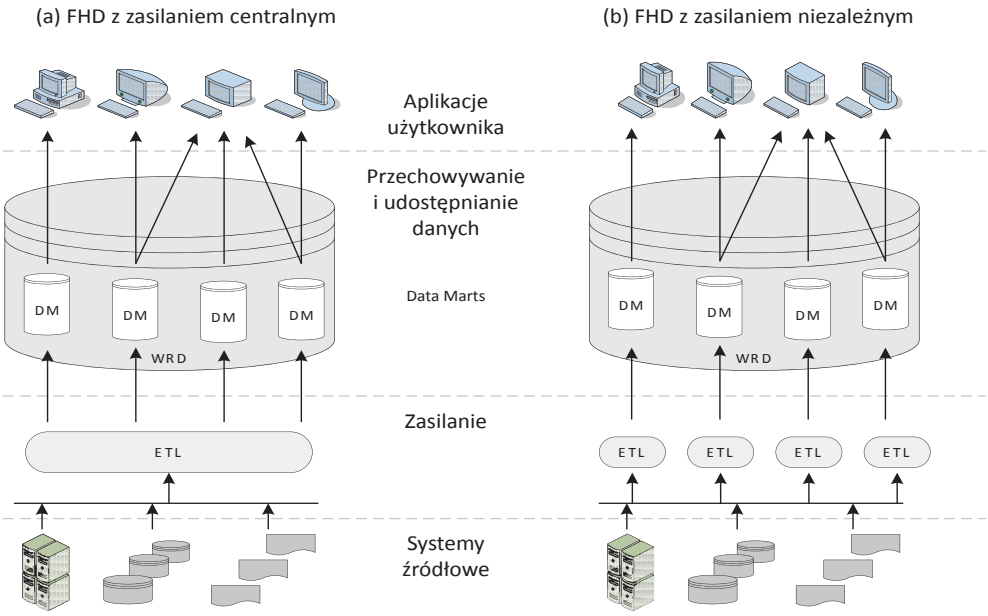
hierarchiczną (rysunek 1, część b), dla podkreślenia większej złożoności architektury w porównaniu z architekturą centralną, również zakłada kopiowanie danych z systemów źródłowych do CRD oraz dodatkowo kopiowanie danych z CRD do mniejszych hurtowni tematycznych, często zintegrowanych z konkretnymi aplikacjami użytkowników. Dzięki takiemu podejściu zostają zachowane zalety architektury centralnej w zakresie jednolitych i zintegrowanych danych, a jednocześnie procesy przetwarzania danych (analizy, raporty itd.) mogą być realizowane w środowisku rozproszonym, nie obciążając bezpośrednio CRD.



Rysunek 1. Warianty korporacyjnej hurtowni danych

Źródło: opracowanie własne.

Oba warianty korporacyjnej hurtowni danych zakładają kopiowanie danych (w wariancie hierarchicznym dwukrotnie), co może rodzić problemy wydajnościowe w przypadku bardzo dużych zbiorów danych lub przy znacznym rozproszeniu geograficznym systemów źródłowych. Między innymi z tych przyczyn opracowano alternatywną architekturę, określaną mianem federacyjnej hurtowni danych, w której odwrócono częściowo kierunek przepływu danych wewnątrz hurtowni, ograniczając równocześnie skalę kopiowania danych. Podobnie jak w przypadku KHD można wyróżnić dwa warianty FHD: wariant ze wspólnymi procesami zasilania (rysunek 2, część a) oraz wariant z niezależnymi procesami zasilania (rysunek 2, część b).



Rysunek 2. Warianty federacyjnej hurtowni danych

Źródło: opracowanie własne.

W podejściu federacyjnym nie występuje centralne repozytorium danych, gdyż dane są umieszczane bezpośrednio w hurtowniach tematycznych. CRD jest zastąpione przez wirtualne repozytorium danych (WRD), które jest jego wirtualną formą. CRD i WRD pełnią podobną funkcję w hurtowni danych, umożliwiając dostęp do wszystkich danych zawartych w hurtowni w jednolitym i zintegrowanym modelu. WRD umożliwia aplikacjom użytkowników korzystanie z danych zlokalizowanych fizycznie w różnych hurtowniach tematycznych, bez konieczności posiadania wiedzy na ten temat. Należy zauważyć, że w wariancie z zasilaniem niezależnym poszczególne hurtownie tematyczne mogą być traktowane jako odrębne korporacyjne hurtownie danych, stąd na gruncie koncepcji federacyjnej hurtowni danych rozważane jest również zagadnienie federacji hurtowni danych[5].

[5] N. Stolba, M. Banek, A.M. Tjoa, *The Security Issue of Federated Data Warehouses in the Area of Evidence-Based Medicine*, First International Conference on Availability, Reliability and Security (ARES '06), 2006, s. 329–339.

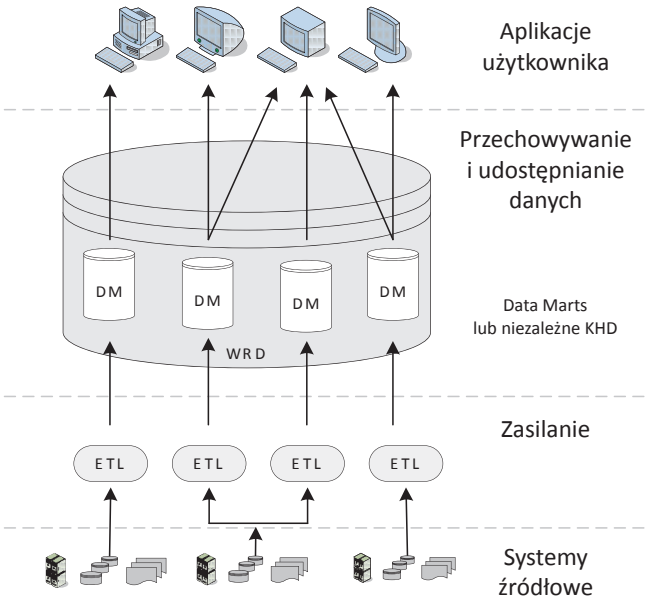
Porównując architekturę KHD i FHD, należy zwrócić uwagę na zalety i wady każdej z nich[6] w ich różnych wariantach, w odniesieniu z jednej strony do zagadnień związanych z efektywnością przetwarzania danych, a z drugiej strony do możliwości skutecznego zarządzania samą hurtownią danych.

Cechą wspólną obu rodzajów hurtowni danych jest istnienie jednolitego wspólnego modelu danych hurtowni, który tworzy płaszczyznę integracji danych i umożliwia dostęp do nich, niezależnie od źródła pochodzenia danych i znajomości modelu danych źródłowych. Dzieję sie tak dzięki istnieniu warstwy zasilającej, w której procesy ETL transformują i integrują dane pochodzące z różnych źródeł. W wariantach współdzielących procesy zasilania (KHD wariant centralny i hierarchiczny oraz FHD wariant z zasilaniem centralnym) implementacja takiego rozwiązania jest stosunkowo prosta, lecz wiąże się z koniecznością dostarczenia danych w jedno miejsce i dopiero ewentualnie późniejszą ich dystrybucją. Aby współdzielenie procesów ETL było możliwe, muszą być dostępne informacje o strukturach danych źródłowych w zakresie niezbędnym do oprogramowania właściwych procesów zasilających. W wariancie FHD z zasilaniem rozproszonym procesy zasilające mogą być implementowane niezależnie. Przy takim rozwiązaniu głównym problemem staje się zapewnienie zgodności lokalnych procesów ETL z modelem danych hurtowni danych. Jakiekolwiek niezgodności praktycznie uniemożliwią powstanie i wykorzystywanie wirtualnego repozytorium danych jako głównej i zintegrowanej bazy danych hurtowni, pozwalającej na współdzielenie danych zlokalizowanych w różnych hurtowniach tematycznych.

Dalszym rozwinięciem wariantu FHD z zasilaniem rozproszonym jest wspomniana wcześniej koncepcja federacji hurtowni danych (rysunek 3). W tej koncepcji poszczególne hurtownie tematyczne mogą być traktowane jako niezależne korporacyjne hurtownie danych z własnymi modelami danych i procesami zasilającymi. Oznacza to brak konieczności przesyłania danych źródłowych oraz upubliczniania informacji o lokalnych strukturach danych – dane mogą być lokalnie transformowane i lokalnie ładowane do struktur hurtowni tematycznych. Informacją upublicznianą w zakresie modelu i struktur danych jest informacja o modelach i strukturach hurtowni tematycznych. Hurtownie tematyczne mogą być budowane na podstawie wspólnego modelu danych WRD, choć również w wariancie bardziej skomplikowanym mogą być oparte na własnych modelach

[6] H.J. Watson, T. Ariyachandra, *Which Data Warehouse Architecture Is Most Successful*, „Business Intelligence Journal", vol. 11, no. 1, s. 4–6, pełny tekst raportu jest dostępny na: www.terry.uga.edu/~hwatson/DW_Architecture_Report.pdf (data odczytu 20.11.2013).

danych, a do stworzenia WRD można wykorzystać mechanizmy zarządzania heterogenicznymi bazami danych. Istotną właściwością federacji hurtowni danych jest daleko idące rozdzielenie systemów źródłowych od wirtualnego repozytorium danych. Właściciel systemu źródłowego może samodzielnie oprogramować procesy ETL i nie musi upubliczniać informacji o wykorzystywanym w nim modelu i strukturach danych. Właściwość ta jest szczególnie istotna w odniesieniu do rozważanego problemu dostępu do danych poufnych, gdy sama informacja o strukturze istniejących baz danych może również mieć charakter poufny.



Rysunek 3. Federacja hurtowni danych

Źródło: opracowanie własne.

3. Uwarunkowania wymiany i dostępu do informacji poufnych

Wspomniane we wstępie organy ochrony prawa podlegają wyjątkowo rygorystycznym regulacjom prawnym związanym z możliwością gromadzenia i wykorzystywania zgromadzonych danych w ramach swojej działalności. Źródłem regulacji prawnych w tym zakresie są:

- ustawy powołujące wspomniane organy i regulujące zasady ich funkcjonowania,
- przepisy wykonawcze do tych ustaw,
- rozporządzenia właściwych organów państwa,
- regulacje prawne o charakterze ogólnym, np. ustawa o ochronie danych osobowych.

Zakres regulacji prawnych obejmuje m.in.:

- przesłanki gromadzenia danych,
- zakres gromadzonych danych,
- możliwość i czas ich przechowywania,
- zasady dostępu do danych,
- możliwość ujawnienia (udostępnienia) danych innym podmiotom (w tym innym organom ochrony prawa).

Dodatkowo na bazie wspomnianych regulacji prawnych każdy z organów opracowuje własne, wewnętrzne regulacje w odniesieniu do poszczególnych zbiorów danych. Celem tak szczegółowych i wielopoziomowych regulacji jest zagwarantowanie, że dane są rzeczywiście gromadzone w związku z realizacją ustawowych obowiązków organu, dostęp do nich jest w pełni kontrolowany i nie zostaną one udostępnione podmiotom nieuprawnionym ani nie będą wykorzystane w sposób niezamierzony przez ustawodawcę. Z drugiej strony, te same regulacje prawne nakładają na poszczególne organy obowiązek wspierania innych organów, również przez udostępnianie danych niezbędnych do realizacji ich zadań[7].

Stawia to organy dysponujące własnymi bazami danych (właścicieli danych) w wyjątkowo trudnej sytuacji. Muszą one zapewnić w maksymalnym stopniu bezpieczeństwo posiadanym danym, a jednocześnie muszą te dane udostępniać w określonym i ograniczonym zakresie. W praktyce oznacza to, że udostępnianie danych w większości przypadków odbywa się dzięki mechanizmom opartym na trybie zapytanie–odpowiedź z wykorzystaniem tradycyjnej formy papierowej. Taki mechanizm wzajemnego udostępniania danych stosunkowo łatwo spełnia

[7] Obowiązek ten wynika nie tylko z krajowych regulacji prawnych, ale również z regulacji prawnych obowiązujących kraje członkowskie UE. Przykładem takiej regulacji może być decyzja ramowa Rady 2006/960/WSiSW z dnia 18 grudnia 2006 r. w sprawie uproszczenia wymiany informacji i danych wywiadowczych między organami ścigania państw członkowskich Unii Europejskiej, zobowiązująca organy ochrony prawa do wzajemnej wymiany danych, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006F0960:PL:HTML (data odczytu 20.11.2013).

wymagania zapewnienia bezpieczeństwa danych, lecz jego niewątpliwą wadą jest zarówno forma, jak i czas potrzebny na udostępnienie danych.

Aby móc zaproponować inne mechanizmy udostępniania danych oparte na technologiach informatycznych, należy określić wymagania, jakie taki system musiałby spełniać. W kontekście bezpieczeństwa danych podstawowym wymaganiem jest pełna kontrola nad dostępem do danych. W szczególności – przy uwzględnieniu różnych poziomów bezpieczeństwa i możliwych zagrożeń – należy przyjąć, że:

1. Struktura i model bazy danych pozostają widoczne jedynie dla właściciela danych, a odbiorca udostępnianych danych nie uzyskuje informacji o sposobie ich przechowywania.
2. Każdy podmiot ubiegający się o dostęp do danych musi być zweryfikowany w zakresie możliwości udostępniania danych.
3. Zakres udostępnianych danych jest weryfikowany w zależności od podmiotu, któremu są one udostępniane.
4. Właściciel danych podejmuje decyzję o udostępnieniu danych podmiotowi, któremu dane są udostępniane, oraz ich zakresie. Decyzja taka może mieć charakter ogólny (tzn. danemu podmiotowi są udostępniane dane o określonym zakresie), lecz może być w każdej chwili zmieniona.
5. Każdy zdarzenie związane z udostępnieniem danych (wystąpienie o dostęp do danych, przekazanie danych, odmowa przekazania danych itd.) jest szczegółowo rejestrowane.
6. Transmisja danych pomiędzy uczestnikami procesu wymiany powinna być realizowana bezpiecznymi kanałami transmisji (np. przy wykorzystaniu szyfrowania i sieci prywatnych). Dane nie powinny być nawet czasowo składowane poza kontrolą uczestników.

Wymaganie 4 nie pozwala na kopiowanie danych, co tym samym wyklucza wykorzystanie systemu opartego na architekturze korporacyjnej hurtowni danych. Analogicznie wymaganie 1 neguje istnienie wspólnej warstwy zasilającej (współdzielenia procesów ETL), co wyklucza również architekturę federacyjnej hurtowni danych w wariancie z zasilaniem centralnym. W tym kontekście jedyną architekturą systemu, jaką można rozważać, jest wariant federacyjnej hurtowni danych z w pełni izolowanymi procesami ETL (federację hurtowni danych). W dalszej części artykułu zostanie pokazana koncepcja wykorzystania mechanizmów hurtowni danych na podstawie architektury FHD do budowy systemu dostępu do danych poufnych, spełniająca wyliczone powyżej wymagania.

4. Bezpieczny dostęp do danych poufnych – schemat procesu

Podstawowym założeniem prezentowanej koncepcji jest wykorzystanie dostępu do danych poufnych w trybie pytanie–odpowiedź, jako trybu z jednej strony zapewniającego wysoki poziom bezpieczeństwa, a z drugiej strony powszechnie wykorzystywanego przez różne organy. W obecnej formie proces dostępu do danych przebiega według następującego schematu:

1. Po identyfikacji potrzeby pozyskania danych mogących się znajdować w zasobach innego organu formułowane jest zapytanie, które najczęściej jest przygotowywane na podstawie gotowego wzorca, wypełnianego jedynie danymi identyfikującymi przedmiot zapytania. Na tym etapie jest już określony potencjalny zakres danych, jakie mogą znaleźć się w odpowiedzi. Zakres ten wynika ze wspomnianych wcześniej regulacji prawnych określających zasady udostępniania danych przez organy ochrony prawa.
2. Zapytanie – po jego akceptacji przez uprawnionego funkcjonariusza organu pytającego – jest przekazywane do pytanego organu, gdzie zapytanie podlega akceptacji przez uprawnionego funkcjonariusza organu pytanego i jest przekazywane do realizacji (w przypadku braku akceptacji do pytającego przesyłana jest odmowna realizacji zapytania).
3. Po przygotowaniu odpowiedzi (czyli wyszukaniu właściwych informacji w bazie danych i nadaniu im odpowiedniej formy) podlega ona akceptacji przez właściwego funkcjonariusza i jest przekazywana do organu pytającego (gdzie może zostać wprowadzona do bazy danych).

Każdy z etapów tego procesu, czyli przygotowanie zapytania, jego akceptacja do wysłania, wysłanie, odebranie przez pytany organ, akceptacja do realizacji, przygotowanie odpowiedzi, jej akceptacja, wysłanie i odbiór przez organ pytający, jest dokumentowany co najmniej w zakresie osób zaangażowanych w realizację poszczególnych działań i czasu ich wykonania.

Z wyróżnionych etapów tego procesu automatyzacji z wykorzystaniem technologii informatycznej może podlegać:

1. Przygotowanie zapytania – wybór i wypełnienie formularza na podstawie wcześniej przygotowanych wzorców zapytań dopuszczalnych pomiędzy danymi organami (rodzaje zapytań).
2. Przygotowanie odpowiedzi – wykorzystanie zestandaryzowanych formularzy pozwala na wykorzystanie przygotowanych wcześniej odpowiednich zapytań do bazy danych uzupełnianych jedynie danymi zawartymi w zapytaniu.

3. Przesłanie zapytania i odpowiedzi – zapytanie i odpowiedź w formie elektronicznej mogą być przesłane siecią komputerową przy zachowaniu odpowiedniego poziomu bezpieczeństwa transmisji danych.
4. Dokumentowanie przebiegu procesu dostępu do danych poufnych – każdy etap tego procesu jest realizowany w interakcji z użytkownikiem, co pozwala na identyfikację osób, ich uprawnień oraz odnotowanie podjętych działań (decyzji) oraz czasu ich podjęcia.

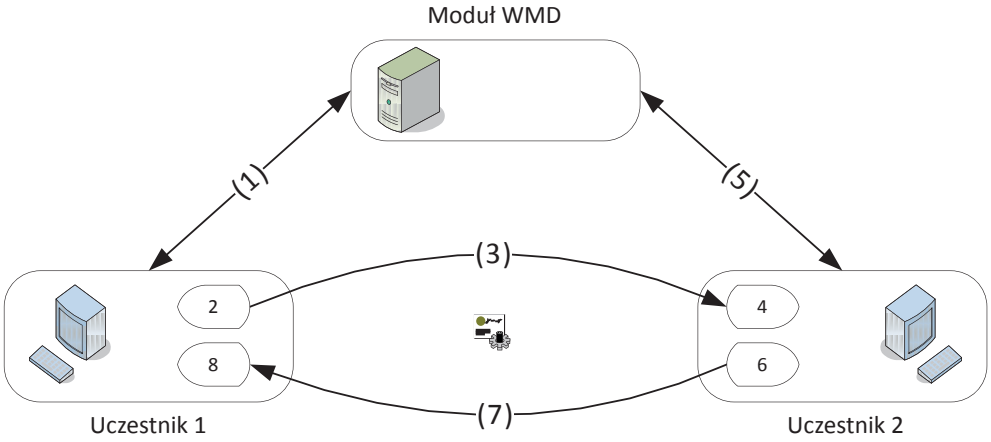
Warto zwrócić uwagę na fakt, że wykorzystanie zestandaryzowanych formularzy zapytań oznacza, że organ pytający nie musi znać modelu i struktur baz danych organu pytającego. Opierając się na predefiniowanych do każdego formularza zapytaniach, można w sposób automatyczny generować właściwe zapytania do bazy danych.

Z odmienną sytuacją mamy do czynienia w przypadku odpowiedzi. Chociaż wynik zapytania nie zawiera wprost informacji o strukturze bazy danych, to pośrednio można uzyskać te informacje, a dodatkowo wykorzystuje on model konkretnej bazy danych. Dlatego odpowiedź przed wysłaniem powinna zostać poddana dodatkowej transformacji, tak aby nie ujawniać tych informacji. Ponadto należy rozważyć możliwy przypadek, gdy dany rodzaj zapytania jest kierowany do więcej niż jednego organu. W takiej sytuacji uzyskane odpowiedzi będą sformułowane w różny sposób (nawet jeśli narzucimy danemu zapytaniu strukturę odpowiedzi, to odpowiedzi będą wykorzystywać różne modele danych), a ich umieszczenie w bazach danych pytającego wymagałoby odrębnych procesów transformacji. Przy dużej liczbie organów udostępniających sobie wzajemnie dane i dużej liczbie rodzajów zapytań powodowałoby to nadmierną komplikację procesów przetwarzania danych.

Rozwiązaniem tego problemu jest opracowanie jednolitego wspólnego modelu danych (WMD) na potrzeby wymiany danych[8]. Możliwość opracowania wspólnego modelu wynika zarówno z charakteru samych danych, jak i z zakresu wymiany. Zakres gromadzonych i udostępnianych wzajemnie przez organy danych wynika z regulacji prawnych. Z przedmiotowego punktu widzenia dane mogą dotyczyć osób, podmiotów prawnych, mienia ruchomego i nieruchomego oraz zdarzeń o potencjalnie przestępczym charakterze. Zakres danych możliwych do gromadzenia i udostępniania w odniesieniu do każdego elementu ma

[8] F. Saltor, M. Oliva et al., *Building secure data warehouse schemas from federated information systems*, w: *Heterogeneous Information Exchange and Organizational Hubs*, red. H. Bestougeff, J.E. Dubois, B. Thuraisingham, Kluwer Academic Publishers, Dorbrecht–Boston–London 2002, s. 123–134.

charakter skończony. Wykorzystując metody modelowania danych w hurtowniach danych, można skonstruować wspólny model danych wystarczający do opisu zasobów danych gromadzonych i udostępnianych wzajemnie przez różne organy[9]. W efekcie tak przygotowany model jest wystarczający do definiowania rodzajów zapytań i do określania formatu odpowiedzi. Dla zachowania bezpieczeństwa danych każdy organ uczestniczący w procesie udostępniania danych (zwany dalej uczestnikiem) ma dostęp do tej części modelu, która opisuje jego zasoby danych. Konsekwencją wprowadzenia WMD jest konieczność zarządzania tym modelem, co oznacza wyróżnienie w architekturze systemu wymiany danych poufnych modułu odpowiedzialnego za utrzymanie i zarządzanie WMD. Schemat procesu dostępu do danych z uwzględnieniem WMD przedstawia rysunek 4.



Rysunek 4. Schemat dostępu do danych z uwzględnieniem WMD

Źródło: opracowanie własne.

Na schemacie zostało wyróżnionych osiem głównych procesów:

1) wybór i pobranie wzorca zapytania opisanego w WMD,
2) wypełnienie danymi zapytania i decyzja o jego wysłaniu,
3) przesłanie bezpiecznym kanałem transmisyjnym zapytania do uczestnika 2,

[9] Prace nad wykorzystaniem m.in. ontologii do stworzenia jednolitej reprezentacji obiektów i zdarzeń opisywanych w bazach danych organów ochrony prawa były prowadzone w ramach projektu INDECT (http://www.indect-project.eu/), por. m.in.: *D6.4 Ontology and Automatic Reasoning In Crisis Management – Definitions and Concepts*, http://www.indect-project.eu/files/deliverables/public/d6.4/at_download/file (data odczytu: 20.11.2013). Opis taki jest naturalnym punktem wyjścia do stworzenia wspólnego modelu danych.

4) odczyt zapytania i decyzja o jego realizacji,
5) pobranie schematu odpowiedzi opisanego w WMD,
6) wykonanie zapytania na bazie danych użytkownika 2, konwersja wyniku do schematu odpowiedzi i decyzja o wysłaniu odpowiedzi,
7) przesłanie bezpiecznym kanałem komunikacji odpowiedzi do uczestnika 1,
8) odczyt odpowiedzi (w modelu WMD), konwersja do lokalnego modelu danych i zapisanie w bazie danych uczestnika 1 (opcjonalnie).

Należy zwrócić uwagę na fakt, że procesy komunikacyjne przesyłające dane poufne, czyli procesy 3 i 7, są realizowane bezpośrednio pomiędzy uczestnikami udostępniania danych. Chociaż przy dużej liczbie uczestników naturalne wydaje się wykorzystanie modułu WMD jako pośrednika we wzajemnej komunikacji, to w kontekście bezpieczeństwa danych takie rozwiązanie jest niedopuszczalne, gdyż dane musiałyby być, choćby czasowo, przechowywane przez ten moduł (np. buforowanie transmisji). Tworzyłoby to potencjalną możliwość dostępu do danych poufnych osobom nieuprawnionym (w prezentowanym schemacie osoby uprawnione do dostępu do danych to jedynie funkcjonariusze uczestników 1 i 2).

Ważną właściwością przedstawionego schematu jest fakt, że informacje o modelu i strukturze danych poszczególnych uczestników są widoczne tylko dla nich samych. Każdy z uczestników indywidualnie tworzy procesy translacyjne do WMD i ma nad nimi pełną kontrolę. Dodatkowo, ponieważ każdy z uczestników jest identyfikowalny w systemie wymiany danych, to jego dostęp do WMD, na poziomie modułu zarządzania WMD, może być ograniczony do zakresu wynikającego z restrykcji prawnych. Oznacza to, że jeżeli z regulacji prawnych wynika, że uczestnik 1 może uzyskać dostęp do danych uczestnika 2, lecz nie może mieć dostępu do danych uczestnika 3, to mimo że uczestnik 2 ma dostęp do danych uczestnika 3, uczestnik 3 pozostaje niewidoczny dla uczestnika 1 (w szczególności nie ma on dostępu do opisu zasobów danych uczestnika 3 w WMD i nie istnieją żadne zdefiniowane wzorce zapytań między nimi). Oznacza to, że przedstawiony schemat dostępu do danych poufnych odzwierciedla bilateralny charakter współpracy między organami ochrony prawa.

Kolejnym ważnym aspektem proponowanego schematu dostępu do danych jest możliwość automatyzacji poszczególnych procesów. W przypadku procesów obejmujących podejmowanie decyzji (składnik procesów 2, 4 i 6) możliwość automatyzacji wynika z przyjętych przez poszczególne organa regulacji wewnętrznych. Przykładowo w przypadku procesu 2, jeśli przyjmiemy, że dostęp do systemu, czyli możliwość formułowania zapytań, mają jedynie

funkcjonariusze, którzy są formalnie uprawnieni do przesłania zapytania do innych organów, zapytanie może być automatycznie przesłane po jego sformułowaniu, bez konieczności odzwierciedlania w systemie podejmowania decyzji i oczekiwania na dodatkowe potwierdzenia. Znacznie istotniejszym z punktu widzenia uczestników i w kontekście bezpieczeństwa danych jest zagadnienie automatyzacji procesu 6 w zakresie wykonania zapytania i przygotowania odpowiedzi. Z technicznego punktu widzenia nic nie stoi na przeszkodzie, aby wykonanie zapytania w pełni zautomatyzować. Z każdym wzorcem zapytania o dane (opisanego w WMD) można powiązać predefiniowane zapytanie do bazy danych (np. skrypt lub zestaw skryptów napisanych w języku zapytań lokalnej bazy danych na podstawie jej modelu i struktury), które zostanie automatycznie wykonane po jego odebraniu, a uzyskane wyniki po automatycznej transformacji do WMD mogą być przesłane jako odpowiedź. Choć taki poziom automatyzacji procesu zapewniłby niewątpliwie najszybsze jego działanie, to niezależnie od kwestii związanych z kontrolą dostępu do danych poufnych (która to kwestia w dużym stopniu stawia takie rozwiązanie pod znakiem zapytania) taki stopień automatyzacji oznaczałby konieczność zbudowania stałego interfejsu pomiędzy systemem wymiany danych a systemem zarządzającym bazą danych danego uczestnika. Niezależnie od wykorzystanych rozwiązań technicznych sam fakt istnienia takiego interfejsu mógłby być traktowany jako istotne zagrożenie nieautoryzowanego dostępu do danych. Z tego też powodu w prezentowanym schemacie wewnętrzne procesy dostępu do baz danych poszczególnych uczestników nie są analizowane, a z założenia w schemacie procesów wymiany danych poufnych występuje podejmowanie decyzji jako istotny element tego procesu.

5. Nadzór nad procesem udostępniania danych poufnych

Zagadnienie udostępniania danych poufnych to zagadnienie dotyczące nie tylko stworzenia bezpiecznego systemu wymiany danych, ale również możliwości kontroli tego, czy dostęp do danych poufnych wynika zawsze z właściwych przesłanek i nie jest nadużywany. Sprawowanie kontroli nad udostępnianiem danych należy rozważać w dwóch kategoriach: kontroli wewnętrznej, sprawowanej przez organ pytający lub pytany, oraz kontroli zewnętrznej, sprawowanej przez inne organy państwa (lub inne podmioty) mające właściwe uprawnienia. Przy czym należy zwrócić uwagę na to, że w przypadku kontroli sprawowanej

przez zewnętrzne organy możliwe są dwa rodzaje kontroli. Część zewnętrznych organów kontrolnych posiada uprawnienia do przeprowadzania kontroli na poziomie wewnętrznym, tzn. z pełnym dostępem do danych. W tym przypadku w kontekście zakresu informacji o wymianie danych poufnych różnica między kontrolą wewnętrzną i zewnętrzną leży głównie w podmiocie prowadzącym kontrolę. Drugi rodzaj kontroli zewnętrznej jest realizowany na podstawie informacji o wymianie danych poufnych udostępnianych na zewnątrz przez same organy uczestniczące w wymianie danych w formie raportów. Ponieważ koncentrujemy się na kwestiach związanych z zakresem informacji o wymianie danych poufnych, w dalszych analizach jako kontrolę zewnętrzną będziemy traktowali jedynie ten drugi rodzaj kontroli. Zestawienie zakresu informacji opisujących pojedynczy proces udostępnienia danych wraz ze wskazaniem zakresu dostępu do tych informacji w ramach czynności kontrolnych zawiera tabela 1.

Tabela 1. Zestawienie informacji o procesie dostępu do danych poufnych wraz ze wskazanie zakresu dostępu do tej informacji w ramach czynności kontrolnych

	Zakres informacji	Organ pytający	Organ pytany	Kontrola zewnętrzna
1.	Organ pytający	tak	tak	tak
2.	Organ pytany	tak	tak	tak
3.	Rodzaj zapytania[10]	tak	tak	częściowo
4.	Kto (dane funkcjonariusza) wystąpił z zapytaniem	tak	nie	nie
5.	W ramach jakich czynności (jakie były przesłanki wystąpienia z zapytaniem)	tak	częściowo	częściowo
6.	Kiedy zapytanie zostało sformułowane	tak	nie	tak
7.	Kto zaakceptował zapytanie	tak	tak	częściowo
8.	Kiedy zapytanie zostało wysłane	tak	tak	tak
9.	Kiedy zapytanie zostało odebrane	tak	tak	tak
10.	Kto zaakceptował wykonanie zapytania	tak	tak	tak
11.	Kiedy zapytanie zostało zaakceptowane do wykonania	tak	tak	tak
12.	Kto wykonał zapytanie i przygotował odpowiedź	nie	tak	nie

[10] Rodzaj zapytania zawiera informacje o zakresie danych, o jakie pytano, oraz o strukturze odpowiedzi opisanej we wspólnym modelu danych (WRD), ale nie zawiera samych danych.

	Zakres informacji	Organ pytający	Organ pytany	Kontrola zewnętrzna
13.	Kiedy odpowiedź była gotowa	nie	tak	nie
14.	Kto zaakceptował odpowiedź	tak	tak	częściowo
15.	Kiedy odpowiedź została wysłana	tak	tak	tak
16.	Kiedy odpowiedź została odebrana	tak	tak	tak
17.	Kto akceptował odebranie odpowiedzi	tak	tak	tak
18.	Kiedy odpowiedź została wprowadzona do bazy danych pytającego	tak	nie	nie
19.	Kto wprowadził odpowiedź do bazy pytającego	tak	nie	nie

tak – posiada dostęp do tej informacji
nie – nie posiada dostępu do tej informacji
częściowo – informacja może być udostępniona, ale nie w pełnym zakresie

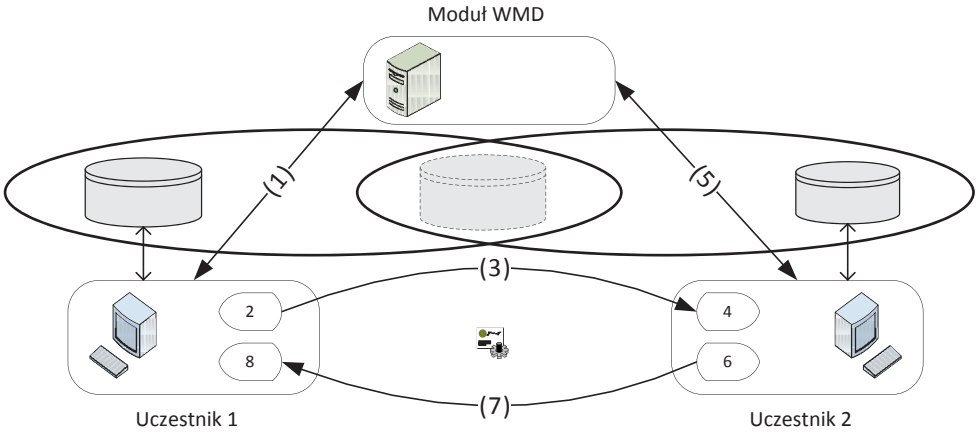
Źródło: opracowanie własne.

Ze względu na fakt, że wskazany zakres informacji opisujących proces udostępniania informacji odzwierciedla przebieg tego procesu z wykorzystaniem zaproponowanego systemu wymiany danych poufnych, informacje te mogą być zbierane automatycznie[11]. Istniejące różnice w zakresie dostępnej informacji o procesie udostępniania danych, zarówno między organami uczestniczącymi w samym procesie udostępniania danych, jak i w przypadku podmiotu zewnętrznego, stanowią istotną przesłankę tego, aby do przechowywania tych informacji i ich analizy wykorzystać architekturę rozproszoną. Na rysunku 5 na schemacie procesu wymiany informacji zaznaczono lokalizację baz danych gromadzących informację opisującą proces wymiany danych.

Informacje o procesie udostępniania danych są gromadzone i przechowywane na poziomie uczestników. Źródłem tych informacji są zarówno działania podejmowane przez funkcjonariuszy uczestników w ramach procesu udostępniania danych, jak i samo zapytanie i odpowiedź. Należy zauważyć, że informacje dostępne do kontroli zewnętrznej stanowią informacje znajdujące się w dyspozycji organu pytającego i organu pytanego, a znakomita większość z nich się pokrywa. Umiejscowienie tych informacji w systemie wymiany danych pozwala na ujednolicenie sposobu ich przechowywania przez uczestników procesu wymiany

[11] W przypadku poz. 18 i 19, przy uwzględnieniu omówionych wcześniej uwarunkowań w zakresie możliwości automatyzacji procesów na poziomie poszczególnych uczestników, informacje te musiałyby być uzupełniane manualnie.

danych. Ponieważ te informacje nie obejmują udostępnianych danych poufnych, a jedynie informacje o fakcie, że do udostępnienia doszło, nie podlegają one tak silnym restrykcjom w zakresie ochrony dostępu do nich.



Rysunek 5. Lokalizacja danych opisujących proces dostępu do danych poufnych
Źródło: opracowanie własne.

W takiej sytuacji można bazy danych gromadzące informacje o udostępnianiu danych potraktować jako źródła zasilania lokalnych hurtowni tematycznych ukierunkowanych na analizę procesów udostępniania własnych danych poufnych lub pobierania takich danych od innych organów przez dany organ. Takie hurtownie tematyczne zawierałyby tylko informacje w zakresie dostępnym do kontroli zewnętrznej. Ponieważ wszystkie hurtownie tematyczne korzystają z jednego modelu danych (który może być traktowany jako część WMD), to jeśli wykorzystałoby się mechanizmy federacyjnej hurtowni danych, dane te mogłyby być integrowane na poziomie modułu zarządzania WMD, tworząc wirtualne repozytorium danych obejmujące informacje na temat wzajemnego udostępniania danych poufnych przez organy ochrony prawa.

Powstała w ten sposób hurtownia danych pozwalałaby m.in. na analizę zakresu częstości i zakresu udostępniania danych czy czasu potrzebnego na realizację różnego rodzaju zapytań. Pozwalałaby również automatycznie generować raporty dla organów nadzorujących.

6. Podsumowanie i kierunki dalszych badań

Mimo potrzeby, jak również w wielu wypadkach obowiązku wzajemnego udostępniania danych poufnych możliwości wspierania tego procesu przez rozwiązania oparte na technologiach informatycznych są bardzo ograniczone. Poufny charakter danych oraz obowiązujące regulacje prawne w zakresie ich ochrony uniemożliwiają integrację danych gromadzonych przez poszczególne służby w jakiejkolwiek formie. Z analogiczną sytuacją mamy do czynienia w zakresie możliwości automatyzacji procesu wzajemnego udostępniania danych.

Przeprowadzona analiza pozwoliła na sformułowanie warunków, jakie musi spełniać system wspierający wzajemne udostępnianie danych poufnych przez organy ochrony prawa. Uwzględniając te warunki, dokonano analizy możliwości wykorzystania mechanizmów opartych na różnych rodzajach architektury hurtowni danych z uwzględnieniem narzędzi zarządzania heterogenicznymi bazami danych. W efekcie przeprowadzonych analiz została wypracowana koncepcja architektury systemu udostępniania danych poufnych przez organy ochrony prawa opierająca się na wykorzystaniu w ograniczonym zakresie jedynie wybranych mechanizmów hurtowni danych.

Wypracowana koncepcja składa się z dwóch głównych elementów – systemu umożliwiającego udostępnianie danych poufnych oraz zintegrowanego z tym systemem rozwiązania wspierającego procesy kontrolne. Zaproponowany system udostępniania danych spełnia zasady bezpieczeństwa wynikające z regulacji prawnych. W swoim działaniu wykorzystuje on w ograniczonym zakresie zaczerpniętą z hurtowni danych koncepcję wspólnego modelu danych do integracji danych. System ten jest uzupełniony o zintegrowany z nim mechanizm budowy hurtowni danych ukierunkowanej na analizę przebiegu procesu udostępniania danych poufnych w zakresie umożliwiającym sprawowanie zewnętrznego nadzoru nad organami ochrony prawa dysponującymi własnymi zasobami danych o charakterze poufnym lub korzystających z zewnętrznych danych.

Uzyskane wyniki stanowią punkt wyjścia do dalszych badań nad udostępnianiem danych poufnych – badań, których celem jest zaprojektowanie i implementacja systemu opartego na przedstawionej koncepcji. Wśród planowanych prac najistotniejsze są kwestie budowy wspólnego modelu danych, w tym wyboru właściwych metod i narzędzi.

Bibliografia

1. Haas L.M., Lin E.T., Roth M.A., *Data integration through database federation*, „IBM Systems Journal" 2002, vol. 41, no. 4.
2. Inmon W.H., *Building the Data Warehouse*, John Wiley & Sons, New York 2005.
3. Saltor F., Oliva M. et al., *Building secure data warehouse schemas from federated information systems*, w: *Heterogeneous Information Exchange and Organizational Hubs*, red. H. Bestougeff, J.E. Dubois, B. Thuraisingham, Kluwer Academic Publishers, Dorbrecht–Boston–London 2002.
4. Stolba N., Banek M., Tjoa A.M., *The Security Issue of Federated Data Warehouses in the Area of Evidence-Based Medicine*, First International Conference on Availability, Reliability and Security (ARES '06), 2006.
5. Watson H.J., Ariyachandra T., *Which Data Warehouse Architecture Is Most Successful*, „Business Intelligence Journal" 2006, vol. 11, no. 1, March.
6. Wrembel R., *Usługi heterogeniczne – techniki integracji rozproszonych baz danych różnych producentów*, Materiały konferencyjne XI Konferencji PLOUG „Systemy informatyczne: projektowanie, implementowanie, eksploatowanie", 18–21 października, Zakopane–Kościelisko 2005.

Źródła sieciowe

1. *D6.4 Ontology and Automatic Reasoning In Crisis Management – Definitions and Concepts*, http://www.indect-project.eu/files/deliverables/public/d6.4/ at_download/file (data odczytu 20.11.2013).
2. *Data Warehouse Architectures: Factors in the Selection Decision and the Success of the Architectures*, www.terry.uga.edu/~hwatson/DW_Architecture_Report.pdf (data odczytu 20.11.2013).
3. Decyzja ramowa Rady 2006/960/WSiSW z dnia 18 grudnia 2006 r. w sprawie uproszczenia wymiany informacji i danych wywiadowczych między organami ścigania państw członkowskich Unii Europejskiej, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006F0960:PL:HTML (data odczytu 20.11.2013).

* * *

Using the concept of federated data warehouse for a safe exchange of confidential information

Summary

Confidential information is collected by various law-enforcement institutions. In order to comply with the regulations and to ensure an effective implementation of their tasks, these institutions have a duty to provide each other with such information to the extent specified by the law. This paper presents a concept of supporting the process of sharing confidential data and supervision of this process using tools taken from the concept of federated data warehouse.

Keywords: data warehouse, confidential information, information exchange