AUTOMATYKA • 2005 • Tom 9 • Zeszyt 3

Wojciech Janicki*

Jakość bazy danych

1. Wprowadzenie

Powszechny rozwój informatyki sprawia, że wkracza ona w coraz to nowe dziedziny życia, systemy informatyczne stają się coraz bardziej rozbudowane, złożone i wyrafinowane. Równocześnie łączenie mniejszych systemów w większe, centralne, zawierające pełniejszą informacje z dziedziny, w której są wykorzystywane, powoduje dynamiczne powiększanie baz danych. Jako jedną z cech bazy danych ostatnio często używa się określenia **jakość bazy danych**.

Powstaje pytanie: co to jest jakość bazy danych.

Na wstępie należy zaznaczyć, iż pojęcie jakości bazy danych może oznaczać zarówno jakość oprogramowania tworzącego bazę [1, 5, 6, 7], czyli jakość narzędzia dostarczanego przez określonego producenta, jak i jakość konkretnej aplikacji stworzonej przy użyciu tych narzędzi, a w szczególności jakość danych zawartych w tej bazie danych.

W pierwszym przypadku można prowadzić rozważania, jakie profesjonalne narzędzie dostępne na rynku można uznawać za najlepsze, czy najlepszym narzędziem jest to, które cieszy się największą popularnością oraz uznaniem użytkowników i ekspertów. Z uwagi na szeroko publikowane testy i rankingi porównawcze konkretnych narzędzi nie będziemy się w niniejszej publikacji zajmować tym aspektem pojęcia jakości bazy danych.

W niniejszym artykule zajmiemy się natomiast jakością bazy danych rozumianą jako jakość zgromadzonych tam informacji i danych [4]. Postaramy się odpowiedzieć na pytania: czy jakość bazy danych może być zła, jakie mogą być przyczyny złej jakości bazy danych i czy możemy poprawić jakość bazy danych, na co ma wpływ jakość bazy danych oraz jakie mogą być skutki złej jakości bazy danych.

2. Pojęcie jakości bazy danych

Jakość bazy danych można najprościej określić, iż jest to jakość zawartej w niej informacji, którą można z niej pozyskać.

* Bank BPH, Kraków

Jako podstawowe cechy decydujące o jakości bazy danych można przyjąć:

- kompletność informacji,
- poprawność informacji,
- spójność danych,
- aktualność,
- dziedzina.

Jako kompletność informacji zawartej w bazie danych należy rozumieć poziom wypełnienia poszczególnych danych (parametrów), można to rozumieć jako liczbę rekordów z wypełnionymi danymi w stosunku do liczby wszystkich rekordów.

Poprawność informacji zawartej w bazie danych to poprawność umieszczonych tam danych, ich zgodność ze stanem faktycznym.

Spójność danych to prawidłowość powiązań pomiędzy parametrami, prawidłowość ich wzajemnych relacji, odniesienia, brak wzajemnego wykluczania się informacji.

Jako aktualność należy rozumieć zawartość w bazie danych zgodnych ze stanem faktycznym, utrzymywanie zgodności z aktualnym stanem faktycznym, pomimo ciągłych zmian w rzeczywistości.

Ostatni wymieniony czynnik – dziedzina – określa obszar informacji do gromadzenia, której przeznaczona baza danych.

3. Zła jakość bazy danych

3.1. Czy jakość bazy danych może być zła

Jak zostało powiedziane, o jakości bazy danych decydują określone czynniki wymienione powyżej. Powstaje pytanie, czy jakość bazy danych może być zła lub dobra [2].

Oczywiste jest, iż baza z częściowo wypełnioną informacją (pustymi rekordami) jest bazą o niskiej jakości. Jednak jakość bazy danych może być różnie interpretowana w zależności od dziedziny gromadzenia informacji, gdyż z jednego punktu widzenia określona informacja jest bardzo istotna – przykładowo nazwisko klienta w bazie klientów, a w innym przypadku może nie mieć żadnego znaczenia – przykładowo nazwisko osoby w bazie do badań socjologicznych. Kompletność informacji musi być ściśle powiązana z dziedziną bazy danych i przeznaczeniem bazy danych.

Analogicznie, analizując poprawność informacji, można zdecydować, iż w przypadku określonej bazy danych mówimy o niskiej bądź złej jakości bazy danych. Wspomniana wyżej baza danych o klientach wypełniona w większości rekordów nazwiskiem-kluczem „Kowalski" (z uwagi na brak informacji o prawidłowych nazwiskach klientów), pomimo iż będzie kompletna, jednak z uwagi na poprawność informacji będzie bazą o niskiej jakości. W przypadku konieczności uzyskania informacji o klientach dla celów korespondencyjnych taka baza będzie nawet bazą danych o ujemnej jakości, gdyż możemy z niej uzyskać informację zafałszowaną. Innym przykładam obrazującym jakość bazy danych z uwagi na poprawność informacji w niej zawartej jest często używany numer identyfikacyjny PESEL czy REGON. Wypełnienie bazy danych fikcyjnymi numerami pomimo spełnienia warunków kompletności analogicznie jak w przypadku nazwiska będzie w przypadku konieczno-

ści odwołania się do tych numerów celem identyfikacji skutkowało brakiem możliwości otrzymania przez użytkownika prawidłowej informacji.

Kolejną cechą charakterystyczną, wymienioną powyżej, jest spójność danych, rozumiana jako prawidłowość powiązań pomiędzy parametrami. Trzymając się wcześniejszych przykładów można zauważyć, iż przechowując w bazie danych numer PESEL i datę urodzenia klienta mamy zależność: sześć pierwszych cyfr numeru PESEL jest zgodna z datą urodzenia, w przeciwnym wypadku nie mamy pewności, który z parametrów jest prawidłowy i wówczas znów mówimy o złej jakości bazy danych.

Oczywiście pomijany tu jest fakt oczywistych błędów z bazie danych związanych z jej uszkodzeniem lub wadliwym działaniem, kiedy to danemu klientowi przypisywany byłby adres czy numer PESEL innego klienta. Tak może się stać w przypadku bazy zawartej w kilku plikach i uszkodzeniu łączących je plików indeksowych. Wówczas też możemy mówić, iż jakość uzyskanej informacji jest zła, jednak nie wynika to wprost z jakości samej bazy danych.

Wspomniane było wcześniej również, iż kolejnym czynnikiem decydującym o jakości bazy danych jest jej aktualność. Baza, która w danym okresie będzie posiadała wysoki poziom jakości, po upływie określonego czasu będzie miała zdecydowanie niższą jakość, jeśli nie zapewnimy sprawnych mechanizmów jej uaktualniania. Tak więc baza nawet o bardzo dobrej jakości danych może stać się bazą o złej jakości danych, gdyż np. klienci zmienią miejsce zamieszkania, stan cywilny, charakter pracy itp., a informacje w bazie nie zostaną uaktualnione.

3.2. Przyczyny złej jakości bazy danych

Powyżej zostały pokazane przykłady złej jakości bazy danych. Należy się teraz zastanowić, co może być przyczyną złej jakości bazy danych [3]. Jak już wspomniano, o tym czy baza jest dobrej, czy złej jakości, decyduje dziedzina, do której przeznaczona jest baza danych. Jeśli posiadamy określoną bazę danych o wysokiej jakości i zdecydujemy się wykorzystać ją w innej dziedzinie, może się okazać, iż będzie to baza o niskiej jakości. Przykładowo, jeśli posiadamy bazę danych określaną jako wysokiej jakości bazę klientów kupujących lody określonego smaku i na tej podstawie wyznaczymy listę najważniejszych klientów, którą zechcemy przenieść do obszaru sprzedaży nawozów sztucznych, to okaże się, iż mamy bazę danych niskiej jakości z powodu zmiany dziedziny. Przeciwnie, ta sama lista klientów przeniesiona do obszaru sprzedaży napojów chłodzących może stanowić bazę danych wysokiej jakości.

Innym powodem złej jakości bazy danych może być omówiony powyżej brak sprawnego modułu i procedur nadzorujących i uaktualniających dane.

Z zagadnieniem uaktualniania bazy danych wiąże się również konieczność zapewnienia jej prawidłowego wypełniania na etapie wprowadzania danych. Musimy wówczas zdecydować, które dane są nam niezbędne, które powinny znaleźć się w bazie danych, a które niekoniecznie muszą być wypełnione. Można wówczas proceduralnie, a jeszcze lepiej – programowo, wymusić na operatorze bazy danych konieczność pozyskania i wprowadzenia określonych danych. Jeśli tego nie zrobimy, będziemy mieli bazę danych złej jakości. Rodzi się tu pewne niebezpieczeństwo, gdyż użytkownik, na którym wymusimy konieczność wprowadzenia określonych parametrów, bez których dane nie zostaną zapisane

do bazy danych, może wprowadzić dane nieprawidłowe, przykładowo fikcyjny numer PESEL. Wówczas, pomimo że będziemy mieli kompletne dane, to będą one nieprawidłowe i bazę musimy uznać za bazę o złej jakości. Można częściowo temu zapobiec, sprawdzając przy zapisie różnego rodzaju warunki, powiązanie między parametrami, jak również zasady i algorytmy danego parametru. Przykładowo, wymieniany już wielokrotnie numer PESEL posiada specjalny algorytm, dzięki któremu możemy określić, czy dany ciąg liczb może stanowić numer PESEL, czy nie.

Inną przyczyną złej jakości bazy danych, może być stała ewolucja systemów informatycznych, co po pierwsze wymusza konwersje pomiędzy kolejno użytkowanymi systemami, a po drugie – występuje niebezpieczeństwo, iż w nowo powstałej bazie danych będzie brakowało parametrów wcześniej niewymagalnych i w tym zakresie baza będzie niekompletna. Konwersje pomiędzy systemami rodzą też niebezpieczeństwo nieprawidłowego „zmapowania" określonych pól występujących w starym i nowym systemie, a czasem jednoznaczne zmapowanie nie jest możliwe, gdy informacja zawarta dotychczas w jednym polu musi zostać podzielona pomiędzy dwa lub więcej pola w nowym systemie. W tym zakresie znajdują się też wszelkiego rodzaju konwersje wynikające z przejęcia bądź łączenia się firm i konieczność ujednolicenia systemów.

Jeszcze inną przyczyną złej jakości bazy danych może być stałe poszerzanie obszarów wspomaganych informatyką, co wiąże się z nowo wprowadzanymi polami do bazy danych i koniecznością ich uzupełnienia dla już występujących rekordów w bazie. Przykładowo firma posiadająca bazę danych klientów, z którymi prowadzi współpracę, zechce lepiej dopasować swoją ofertę do oczekiwań poszczególnych klientów. W tym celu w bazie danych zostaną zaimplementowane dodatkowe pola do przechowywania nowej, dodatkowej informacji. Jednak już występujący w bazie danych klienci nie posiadają przypisanej tej informacji, co powoduje, iż na tym etapie baza danych jest złej jakości, a jakość ta będzie z czasem dopiero podnoszona.

Baza danych może też stać się bazą danych o złej jakości wskutek błędów przetwarzania (błędy systemowe), które niezauważone w porę mogą zafałszować informację w niej zawartą. Jako szczególny rodzaj tej grupy błędów należy zaznaczyć błędy związane z przesyłaniem informacji, gdyż obecnie coraz więcej systemów to systemy centralne obsługujące wiele oddziałów danej firmy.

Podsumowując rozważania na temat przyczyn złej jakości bazy danych, możemy podkreślić następujące jej przyczyny:

- brak kompletności na etapie wprowadzania,
- brak poprawności na etapie wprowadzania,
- brak uaktualnień,
- niedopasowanie dziedziny,
- ewolucja systemów informatycznych,
- konwersje pomiędzy systemami informatycznymi,
- rozszerzenie zastosowania informacji zawartej w bazie danych,
- błędy przetwarzania.

Jak widać, wiele czynników jest odpowiedzialnych za fakt, iż jakość określonej bazy danych jest zła, bądź baza o wysokiej jakości może się okazać bazą o złej jakości danych.

3.3. Czy możemy podnieść jakość bazy danych

W dalszej kolejności będzie rozważana jakość bazy danych w aspekcie jej wpływu na różne czynniki. Najpierw należy się zastanowić, czy możemy poprawić jakość bazy danych [2, 3, 4]. Skoro baza o wysokiej jakości może się stać bazą o złej jakości, to również powinniśmy mieć możliwość poprawienia jakości bazy danych. Dla oceny jakości bazy danych konieczne są określone kontrole bazy. Kontrole takie powinny być prowadzone okresowo i objąć okresowym monitoringiem bazę danych, dzięki czemu będzie możliwe bardzo szybkie zauważenie spadku jakości danych i podjęcie czynności, aby temu przeciwdziałać.

Na podstawie uzyskiwanych różnego rodzaju raportów i logów błędów stwierdzających nieprawidłowości w bazie danych możemy:

- zapewnić uszczelnienie systemu na etapie wprowadzania danych tak, aby operator nie mógł pozostawić niewypełnionych pól bądź wprowadzić nieprawidłowe wartości;
- wdrożyć algorytmy i procedury uaktualniania danych;
- przeprowadzić akcje uzupełniania danych wcześniej niepozyskiwanych od klientów;
- wdrożyć procedury szybkiego identyfikowania przypadków błędnego przetwarzania danych.

Osobnym problemem jest zła jakość bazy danych wynikła z konwersji z innych systemów. W tym przypadku w ramach przeprowadzania konwersji należy zapewnić szczegółowe procedury weryfikacji prawidłowości konwersji, jak również wskazane byłoby przeprowadzenie próbnych konwersji, lub wcześniejszej konwersji na próbce danych.

3.4. Na co ma wpływ zła jakość bazy danych

Jak zostało wyżej pokazane, baza danych może być dobrej albo złej jakości, należy się teraz zastanowić, czy i jakie znaczenie może mieć zła jakość bazy danych [2]. Okazuje się, iż jakość bazy danych ma coraz większe znaczenie i staje się wręcz kluczowym problemem, z uwagi na ilość zawartych w bazach informacji i możliwości ich wykorzystania.

Kluczowe obszary, na które ma wpływ jakość bazy danych, to:

- prawidłowość identyfikacji i obsługi klienta,
- prawidłowość przetwarzania danych,
- prawidłowość sporządzania danych sprawozdawczych,
- prawidłowość danych zarządczych,
- prawidłowość działań marketingowych.

Jako pierwszy z wymienionych wyżej obszarów, na który ma wpływ jakość bazy danych, został określony obszar identyfikacji i obsługi klienta. Jest to niezwykle ważny obszar, gdyż aby zapewnić prawidłowość transakcji, klient musi zostać prawidłowo zidentyfikowany, gdyż nie można dopuścić do tego, aby płatność bądź towar zostały przekazane do innego klienta, który co prawda posiada identyczną nazwę, lecz inny adres. Jeśli nie będzie zapewniona odpowiednia jakość danych w zakresie identyfikacji klienta (dane adresowe, numer PESEL/REGON, data urodzenie klienta itp.), istnieje bardzo duże potencjalne ryzyko dokonania transakcji z niewłaściwym klientem. Tak samo prawidłowość innych parame-

trów zawartych w bazie danych ma wpływ na prawidłowość transakcji, jak na przykład karne stopy procentowe, wysokość rabatu, przypisanie klienta do określonej grupy marketingowej, może istotnie zaważyć na przeprowadzanej transakcji. Tak więc niezapewnienie wysokiej jakości bazy danych w tym obszarze, rodzi wysokie ryzyko operacyjne i stawia pod znakiem zapytania wiarygodność i stabilność firmy.

Wpływ jakości bazy danych na przetwarzanie danych w systemie informatycznym jest oczywisty i wynika wprost z przeznaczenia bazy danych. Każda baza danych zaimplementowana w określonym systemie służy do gromadzenia danych tak, aby można było w szybki sposób uzyskać określoną informację, w przypadku bazy danych klientów jest to np. informacja o adresie czy numerze identyfikacyjnym klienta, w przypadku usług finansowych wykonywane są codzienne operacje przetwarzania danych w celu ustawienia prawidłowych stanów na poszczególnych rachunkach i kontach. Zła jakość danych w tym zakresie może spowodować nieprawidłowe wyniki przetwarzania, problemy z przetworzeniem danych, a nawet uniemożliwić przetworzenie danych w określonych przypadkach.

Istotnym problemem jest zapewnienie odpowiedniej jakości danych w celu sporządzania różnego rodzaju sprawozdań. Każda firma, każda instytucja ma obowiązek sporządzania w określonych terminach różnego rodzaju dla różnych instytucji nadzorczych i nadrzędnych. Sporządzanie tych sprawozdań ma na celu stałą kontrolę prawidłowości funkcjonowania firmy, prawidłowości dokonywanych transakcji, rzetelności w przekazywaniu różnego rodzaju płatności na rzecz państwa, instytucji samorządowych itp. Rzetelność tych sprawozdań zależy wprost od jakości bazy danych, gdyż sprawozdania sporządzone ze złej jakości danych będą złe. Istotność sprawozdań to nie tylko nadzór nad firmami, tego typu nadzór ma na celu również zapewnienie stabilności firmy. W bankach obecnie trwają gorączkowe przygotowania do zapewnienia sprawozdawczości zgodnej z Nową Umową Kapitałową (znanej też jako Basel II), która stanowi zalecenia Komitetu Bazylejskiego dla banków odnośnie do zapewnienia odpowiedniego poziomu kapitału w stosunku do udzielanych kredytów. Wiadomo jest, iż z udzielaniem kredytów wiąże się określone ryzyko, w związku z czym każdy bank musi z określonym prawdopodobieństwem być przygotowany na niespłacenie zobowiązania przez dłużnika i nieodzyskanie pożyczonego kapitału. W tym celu Nowa Umowa Kapitałowa doprecyzowuje metody wyliczania ryzyka o nowe, dodatkowe rodzaje ryzyka, jak również stwarza możliwości bardziej indywidualnego i lepiej dopasowanego do potrzeb danej instytucji wyliczania ryzyka i wysokości wymaganego kapitału, aby zapewnić stabilność kapitałową. Aby jednak nowe metody prawidłowo działały i przynosiły odpowiednie korzyści dla banków, konieczne jest zapewnienie wysokiej jakości bazy danych, gdyż tylko w przypadku wyliczeń opartych na rzetelnych i kompletnych danych może być mowa o prawidłowości nadzoru. Analogiczny wpływ i zależności występują we wszelkiego innego typu instytucjach.

Analogicznie do powyższego wpływu na rzetelność i adekwatność sprawozdań jakość bazy danych ma też wpływ na prawidłowość danych zarządczych. Zarządzanie dużymi, ale i również małymi firmami prowadzone jest na podstawie przesłanek płynących z posiadanej bazy danych. Trafność podejmowanych decyzji jest więc wprost od niej uzależniona i bardzo ściśle powiązana. Decyzje organów zarządzających firmą podejmowane na pod-

stawie nieprawidłowych raportów często są błędne, a obecnie trudno sobie wyobrazić większą firmę bez sprawnie działającego informatycznego systemu wspomagającego zarządzanie. Wynika to przede wszystkim z wielkości gromadzonych i mających wpływ na podejmowane decyzje danych, jak również z wymogu rynku, który wymaga bardzo szybkich decyzji; na ich przeanalizowanie w sposób tradycyjny często nie ma czasu. Tak więc, aby można było podejmować prawidłowe decyzje, muszą one być wsparte rzetelnymi danymi w bazie. Trudno na przykład podejmować decyzje o strategii, mając dane dotyczące dochodowości jedynie 20–30 procent klientów lub jeśli posiadamy informacje o dochodowości tych klientów sprzed np. pięciu lat. Zapewnienie prawidłowej informacji zarządczej jest szczególnie ważne w obecnym czasie łączenia firm, powstawania ogromnych koncernów i grup kapitałowych.

Kolejnym wspomnianym obszarem, na który ma wpływ jakość bazy danych, jest obszar marketingu i reklamy, obszar o szczególnie dużej dynamice rozwoju w ostatnim okresie. Firmy w celu pozyskania nowych klientów lub sprzedaży nowych produktów klientom muszą bardzo dobrze poznać ich upodobania. Aby przeprowadzić udaną kampanię reklamową, muszą z dużym prawdopodobieństwem określić, iż trafi ona do właściwych adresatów. Jest to szczególnie ważne z uwagi na bardzo wysokie koszty ponoszonej działalności marketingowej, a reklamowanie przykładowo nowego rodzaju systemów ogrzewania w okolicach równika stanowiłoby tylko stracone koszty. Jednak aby właściwie dopasować akcję marketingową w tym przypadku, musimy posiadać w bazie danych prawidłową informację dotyczącą strefy zamieszkania klienta. Działy marketingu poszczególnych firm wymagają obecnie coraz to nowych informacji pozyskiwanych od klienta, które mają służyć prawidłowości tych działań, stworzenia dedykowanych produktów dla określonej grupy klientów oraz zorganizowania kampanii reklamowej, która dotrze jedynie do tej ściśle określonej grupy klientów. Problemem jest jednak brak szeregu tego typu informacji w przypadku klientów znajdujących się wcześniej w bazie danych. I w tym przypadku, aby skuteczność działań marketingowych była wysoka, konieczne jest podnoszenie i zapewnienie wysokiego poziomu jakości bazy danych.

4. Skutki złej jakości bazy danych

Na zakończenie rozważań nad jakością bazy danych należy się zastanowić, jakie mogą być skutki niezapewnienia odpowiednio wysokiego poziomu jakości bazy danych. Wiele bezpośrednich skutków zostało już omówionych. Przypomnijmy, iż zła jakość danych skutkuje nieprawidłową obsługą klientów, nieprawidłową informacją sprawozdawczą i zarządczą, jak również mało skutecznymi działaniami marketingowymi.

Można też stwierdzić, iż niewłaściwa jakość danych ma odzwierciedlenie w każdym obszarze działania firmy czy instytucji, co może rodzić poważne konsekwencje finansowe, prawne, zagrażać stabilności i płynności firmy, a w skrajnych przypadkach spowodować upadek firmy.

Reasumując, można stwierdzić, iż zapewnienie dobrej jakości bazy danych, stały monitoring i dążenie do podnoszenia jakości bazy danych powinny być zadaniem o jednym z najwyższych priorytetów w firmie czy instytucji.

Literatura

[1] Connolly T., Begg C.: *Systemy baz danych – projektowanie, wdrażanie i zarządzanie w praktyce. Tom 1,2*. Warszawa, ReadMe 2004
[2] English L.P.: *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. New York, John Wiley & Sons Inc. 1999
[3] Gancarz Ł: *Data Quality – kompleksowa metodologia i narzędzia do zapewnienia wysokiej jakości danych w systemach informatycznych*. Bratysława, SAS Forum, SAS Institute, 2004
[4] Olson J.: *Data Quality*. San Francisco, Morgan Kaufmann Publishers 2003
[5] Robert J.M.: *Bazy danych. Język UML w modelowaniu danych*. Warszawa, Mikom 2000
[6] Stanik J., Kwiatkowski P.: *Zapewnienie jakości systemów informatycznych – koncepcja zapewnienia jakości*. WAT, 2000
[7] Stanik J., Kwiatkowski P.: *Zapewnienie jakości systemów informatycznych – elementy dobrej praktyki*. WAT, 2000