Marcin Mazurek

Wydział Cybernetyki
Wojskowa Akademia Techniczna

Architektura systemu wspomagania decyzji medycznych wykorzystująca technologię przetwarzania danych *big data*

1. Wstęp

Rozwój technologii przetwarzania danych może przynieść wymierne korzyści w służbie zdrowia. Zastosowanie nowoczesnych urządzeń diagnostycznych oraz zautomatyzowanych systemów monitorowania zdrowia pacjenta powoduje, że z każdym przypadkiem medycznym jest wytwarzany duży wolumen danych opisujący historię leczenia i stan pacjenta. Dane te razem z towarzyszącymi im diagnozami stawianymi przez lekarzy mogłyby stanowić cenne źródło wiedzy w procesie stawiania diagnozy, określania czynników ryzyka czy też podejmowania decyzji o sposobie leczenia. Dostęp do centralnego repozytorium analitycznego, w którym można odnaleźć zapisy dotyczące innych przypadków medycznych, jest sposobem na dystrybucję wiedzy i doświadczeń, prowadzących do zwiększenia jakości procesu leczenia[1]. Kolejnym krokiem jest analiza tych danych umożliwiająca wykrycie wzorców, reguł klasyfikacyjnych i asocjacji.

Z technologicznego punktu widzenia głównym problemem związanym z wdrożeniem tego typu rozwiązań jest wolumen i format gromadzonych danych. Szeregi czasowe, obrazy, sekwencje wideo udostępniane przez aparaturę diagnostyczną wymagają dużej przestrzeni dyskowej, a możliwości ich analizy z zastosowaniem systemów hurtowni danych oraz dostępnych narzędzi *Business*

[1] N. Savage, *Better Medicine through Machine Learning*, „Communications of the ACM" 2012, vol. 55, no. 1, s. 17–19.

Intelligence są bardzo ograniczone. Podobnie sytuacja wygląda w przypadku opisów tekstowych wprowadzanych przez lekarzy do karty choroby – czy to w formie papierowej, czy elektronicznej. Nieustrukturalizowane dane wymagają dużych nakładów w zakresie sprzętu, oprogramowania, znajomości metod analizy danych.

Wyzwaniem projektowym jest stworzenie systemu analitycznego, który nie tylko służy analizie eksploracyjnej danych historycznych, ale również mógłby być wykorzystywany w bieżącej pracy lekarzy do wsparcia diagnostyki.

Proponowanym sposobem rozwiązania tego problemu jest wykorzystanie możliwości, jakie stwarza technologia przetwarzania dużych, nieustrukturalizowanych danych, określana jako *big data*. Artykuł przedstawia koncepcję zintegrowania zasilonego on-line repozytorium słabo ustrukturalizowanych danych z hurtownią danych opartą na strukturach relacyjnych lub wielowymiarowych. Opisywana architektura systemu ma zapełnić lukę pomiędzy gromadzonymi opisami przypadków medycznych a wykorzystaniem ich w procesie leczenia i badaniach naukowych lekarzy. System zapewni dostęp do jak najszerszej bazy przypadków medycznych oraz da możliwość tworzenia i udostępniania różnego rodzaju modeli analitycznych szerokiej grupie użytkowników. Jako modele analityczne są rozumiane różnego rodzaju procedury obliczeniowe, wyindukowane w procesie analizy danych historycznych, które mogą następnie służyć do oceny nowych przypadków medycznych. Najszersze zastosowanie znajdują modele klasyfikacyjne, pozwalające na określenie przynależności obiektu do jednej z wyróżnionych wcześniej klas na podstawie cech tego obiektu. W przypadku zastosowań medycznych badanym obiektem będzie pacjent lub opis przypadku medycznego (hospitalizacji), wśród cech znajdą się wyniki badań, cechy fizjologiczne, cechy opisujące warunki bytowe itd. Wynikiem klasyfikacji może być wskazanie jednostki chorobowej (diagnoza) czy też najlepszego sposobu leczenia.

Układ artykułu jest następujący: część 2 przedstawia najważniejsze założenia projektowanego systemu, które determinują w największym stopniu jego architekturę. Następnie przedstawiono najważniejsze założenia technologii *big data*, która zostanie uwzględniona w zaprezentowanej w kolejnej części architekturze systemu. Część 5 opisuje koncepcję wdrożenia systemu, w którym usługi analityczne są udostępniane w chmurze. Artykuł zamyka podsumowanie, w którym zebrano również uwagi dotyczące możliwości wdrożenia rozwiązania.

2. Analiza uwarunkowań

Punktem wyjścia do analizy wymagań jest określenie użytkowników systemu oraz scenariuszy wykorzystania systemu. Głównymi grupami użytkowników systemu analitycznego będą:

- Lekarze naukowcy, którzy będą budowali modele analityczne, oceniali ich jakość oraz udostępniali je lekarzom. Powinni dysponować narzędziem umożliwiającym realizację tych zadań bez programowania na podstawie przygotowanych struktur danych, z których wskazywaliby interesujące obiekty.
- Lekarze prowadzący leczenie – odbiorcy zbudowanych modeli diagnostycznych (klasyfikujących) i grupujących (wyszukiwanie podobnych przypadków). Sposób komunikacji z systemem powinien być nastawiony na efektywność wprowadzania danych wejściowych do analizy oraz czytelną formę prezentacji wyników, które będą pomocne w bieżącym podejmowaniu decyzji podczas wizyty ambulatoryjnej czy przy łóżku chorego. Możliwość przeprowadzania w trybie on-line wnioskowania dla konkretnego przypadku wymusza ładowanie na bieżąco danych do systemu analitycznego. Tymi danymi są zarówno wyniki badań, jak i dane wprowadzane przez lekarza do karty choroby.
- Analitycy danych (ang. *data-scientist*), którzy na podstawie kompletnego wolumenu danych zgromadzonych w bazie NoSQL będą w stanie opracowywać algorytmy ekstrakcji interesujących informacji oraz wzorców. Będą oni dysponowali na wejściu znacznie szerszą bazą danych nieustrukturalizowanych niż grupa lekarzy naukowców, jednakże do ich analizy będą musieli sięgnąć po narzędzia wymagające programowania i znajomości systemów zarządzania bazami danych.

Przechowywane w repozytoriach dane powinny zostać odpersonalizowane, tj. pozbawione cech i informacji, które pozwolą na powiązanie danych z konkretną osobą. Dotyczy to przede wszystkim danych pacjentów, ale również lekarzy prowadzących. Jednocześnie uprawnieni lekarze powinni mieć możliwość prezentacji wyników dla „własnych" przypadków chorobowych.

Wsparcie lekarzy naukowców w zakresie automatycznego pobierania wskazanych obiektów z relacyjnego repozytorium analitycznego wymaga uwzględnienia złożoności wynikającej z możliwych wariantów realizacji modelu danych:

- Rozproszenie wyników badań w dużej liczbie tabel, z których każda posiada odmienną strukturę. W sytuacji, gdy każdemu z typów badania diagnostycznego odpowiada oddzielna tabela, łączenie tych danych w celu otrzymania

widoku, w którym jednemu przypadkowi medycznemu odpowiada jeden wiersz, jest problematyczne.

- Załadowanie wszystkich wyników do pojedynczej tabeli, w której każdemu przypadkowi odpowiada jeden wiersz (taki model danych jest obsługiwany przez narzędzia eksploracji danych), wiąże się z następującymi problemami:
  a) Duża liczba kolumn wynikowego zbioru danych powoduje, że macierz jest rzadka, a wartości puste (dla badań, które nie były zlecone pacjentowi) nie mogą być imputowane na podstawie danych z badanej populacji.
  b) Wyniki tych samych badań przechowywane dla różnych przypadków medycznych nie są porównywalne ze względu na różną chwilę wykonania badań w odniesieniu do stadium rozwoju choroby.
  c) Jeżeli badania są powtarzane wielokrotnie, należy wyznaczyć metodę ich agregacji lub dopuścić do zmiennej liczby kolumn wektora wejściowego.
  d) Współwystępowanie u pacjenta jednostek chorobowych i brak możliwości separacji wyników badań związanych z każdą z jednostek chorobowych.
  e) Wyniki badań są wprowadzane w postaci opisów (tekst) przygotowywanych przez lekarzy z użyciem żargonu i skrótów. Dla wyekstrahowanej zawartości tych pól nie da się zaprojektować statycznej struktury, która mogłaby je przechowywać.

Przedstawiona powyżej specyfika danych medycznych w powiązaniu z relacyjnym modelem danych wykorzystywanym przez narzędzia eksploracji danych powoduje, że aby zwiększyć potencjał analityczny rozwiązania, należy wykorzystać możliwości, jakie daje technologia przetwarzania danych *big data*, opisana w kolejnej części artykułu.

3. Założenia koncepcji *big data*

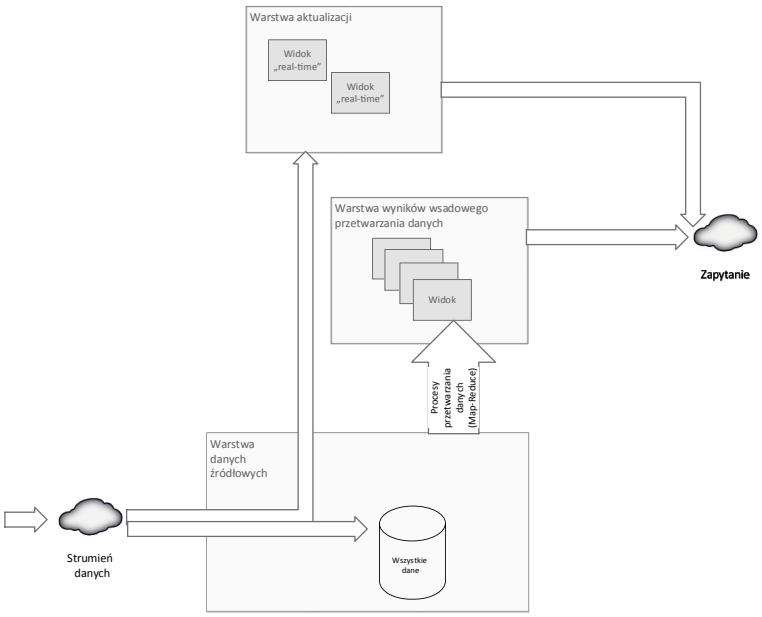
Big data to technologia przetwarzania dużych zbiorów danych, które ze względu na wolumen, szybkość napływu oraz brak struktury nie mogą być wydajnie przetwarzanie w systemach relacyjnych. Najważniejszymi wyróżnikami tej technologii są:

- Rozproszenie danych na wielu węzłach.
- Nierelacyjny system zarządzania rozproszoną bazą danych (NoSQL).

- Rozproszenie przetwarzania danych pomiędzy wiele węzłów na podstawie paradygmatu programowania *map-reduce*.

Referencyjna architektura przetwarzania danych w systemach *big data* odbywa się na podstawie trzech warstw[2] (przedstawionych na rysunku 1):

- Warstwa danych źródłowych – procesy w tej warstwie mają służyć ładowaniu danych ze źródeł do rozproszonego repozytorium. Odmiennie niż w przypadku procesów ETL w hurtowni danych, dane nie są w żaden sposób transformowane – trafiają do repozytorium w postaci źródłowej.
- Warstwa wyników wsadowego przetwarzania danych – jej zadaniem jest przetworzenie zgromadzonych danych zgodnie z algorytmami analitycznymi wyspecyfikowanymi przez użytkownika. Wyniki przetwarzania w stosunku do danych źródłowych charakteryzują się opóźnieniem wynikającym ze wsadowego charakteru procesu.
- Warstwa aktualizacji – uzupełnia wyniki wsadowego przetwarzania danych o dane napływające on-line.



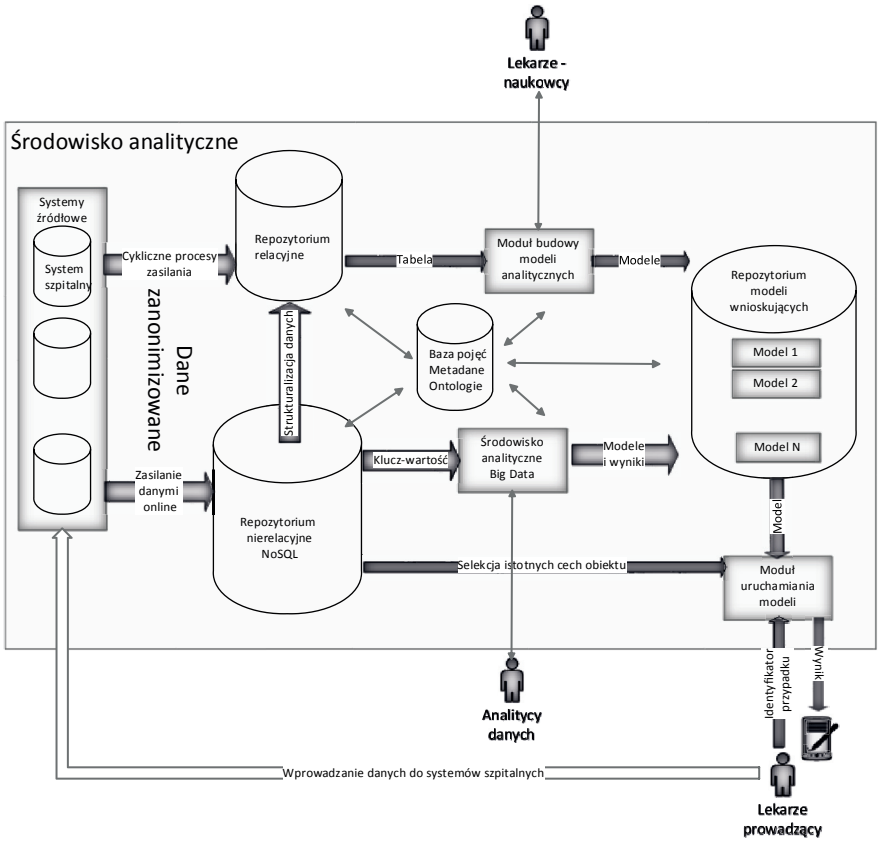
Rysunek 1. Architektura przetwarzania danych w systemie *big data*

Źródło: opracowanie własne na podstawie: N. Marz, *Big Data Principles and best practices of scalable realtime data systems*, MEAP Edition, Manning Publications, Greenwich, CT, USA 2012.

[2] N. Marz, *Big Data Principles and best practices of scalable realtime data systems*, MEAP Edition, Manning Publications, Greenwich, CT, USA 2012.

4. Architektura techniczna rozwiązania

Najważniejszym założeniem architektury jest współwystępowanie repozytorium relacyjnego oraz repozytorium nierelacyjnego (bazy NoSQL), zintegrowanych semantycznie na podstawie ontologii i jej odwzorowania w struktury danych obydwu repozytoriów. Występowanie w architekturze pierwszego z repozytoriów, zawierającego wybrany podzbiór danych zgromadzonych w repozytorium nierelacyjnym, jest podyktowane wymaganiami istniejących narzędzi eksploracji danych. Baza danych NoSQL przechowuje zgromadzone dane w strukturach klucz–wartość, pozwalających zgromadzić wszystkie dane dotyczące przypadków medycznych, nawet jeżeli zakres informacyjny, format lub krotność odbiega od założonego początkowo modelu relacyjnego. Zarys architektury przedstawiono na rysunku 2.



Rysunek 2. Architektura systemu analitycznego danych medycznych
Źródło: opracowanie własne.

4.1. Źródła danych i cykl zasilania hurtowni

Źródłami danych w systemie są: systemy szpitalne, urządzenia diagnostyczne, rejestry leków, dane wprowadzane przez lekarze w trakcie wywiadów. Dane te w postaci źródłowej trafiają do repozytoriów systemu. Można wyróżnić następujące ścieżki ładowania danych:

- Cykliczne ładowanie danych – dla ustrukturalizowanych danych wolnozmiennych. Tą ścieżką mogą być ładowane np. dane finansowe i słowniki procedur.
- Ładowanie danych on-line – pobierane na bieżąco z systemów źródłowych w trybie *pull* bądź *push*, w zależności od możliwości urządzeń i systemów będących źródłem danych, np. różnego rodzaju platform integracyjnych[3]. Ten tryb ładowania danych zapewnia, że w chwili, gdy lekarz wprowadzi dane do systemu szpitalnego, będą one dostępne również w systemie analitycznym i będą mogły zostać poddane analizie z zastosowaniem udostępnionych modeli. Dane te trafiają do repozytorium nierelacyjnego.
- Ładowanie ustrukturalizowanej zawartości z repozytorium NoSQL do bazy relacyjnej. Ekstrakcja informacji odbywa się na podstawie zdefiniowanych przez analityków danych algorytmów wydobywania danych z pól tekstowych, obrazów i sekwencji wideo.
- Wzbogacanie danych źródłowych rezultatami wykonanych analiz. Przykładem takiej zwrotnej danej jest wyznaczony identyfikator skupienia w przeprowadzonej analizie skupień lub wyestymowane wartości parametrów, których brakowało w zbiorze wejściowym.

Procesy zasilające muszą zawierać logikę odpersonalizowania ładowanych danych, z jednoczesnym zachowaniem identyfikatorów, umożliwiających lekarzom prowadzącym zapytanie o dane konkretnego pacjenta lub przypadku medycznego (ang. *reversible coding*)[4].

4.2. Repozytorium NoSQL

Głównym repozytorium infrastruktury jest repozytorium odpersonalizowanych przypadków medycznych, które ze względu na charakter napływających danych można zaklasyfikować jako *big data* (obecność danych obrazowych,

[3] T. Górski, *Architektura platformy integracyjnej dla elektronicznego obiegu recept*, „Roczniki" KAE, z. 25, Oficyna Wydawnicza SGH, Warszawa 2012, s. 67–83.

[4] K.E. Emam, *Guide to the De-Identification of Personal Health Information*, CRC Press, Boca Raton 2010.

opisów tekstowych). Obecność tego typu komponentu w architekturze rozwiązania jest podyktowana następującymi względami:

- Łatwość implementacji i efektywność procesu transferu danych z systemów źródłowych w trybie on-line. Dane w tej bazie pojawiają się natychmiast po wprowadzeniu danych do systemu szpitalnego. Lekarz wprowadza dane raz i są one od razu dostępne jako obiekt analizy.
- Łatwość pobierania danych według klucza pacjenta i nazwy cechy, bez konieczności znajomości struktury danych. W przypadku, gdy wybrane cechy pacjenta stanowią wejście (są zmiennymi objaśniającymi) w modelu klasyfikującym, interfejs bazy klasy NoSQL umożliwia pobranie odpowiednich danych przy użyciu wyłącznie nazw cechy (kluczy) oraz identyfikatora pacjenta. Jako klucze cech będą używane pojęcia ze zdefiniowanego słownika pojęć medycznych (ontologii).
- Możliwość przechowywania danych nieustrukturalizowanych: obrazów, zapisów źródłowych sygnałów, szeregów czasowych, opisów i innych.
- Skalowalność – możliwość wielokrotnego zwiększania pojemności przez dokładanie zasobów sprzętowych bez spadku efektywności zapytań, dzięki zastosowaniu wbudowanej infrastruktury zrównoleglania przetwarzania (*map-reduce*).

4.3. Repozytorium relacyjne

Repozytorium relacyjne przechowuje ustrukturalizowany obraz przypadku medycznego. Składają się na niego:

a) Czynniki ryzyka i dane demograficzne (wiek, płeć).
b) Wyniki badań diagnostycznych i objawy na początku leczenia, w trakcie i po jego zakończeniu.
c) Rozpoznane jednostki chorobowe.
d) Zastosowane procedury medyczne.

Dane te mogą być pobierane bezpośrednio z systemów źródłowych bądź z repozytorium NoSQL.

Istniejące środowiska budowy modeli eksploracji danych zakładają, że ciąg uczący stanowią rekordy zapisane w jednej tabeli relacyjnej, w której każdy z wierszy reprezentuje dokładnie jeden obiekt będący przedmiotem analizy (każdy rekord przechowuje dane dla jednego pacjenta). Pierwszy etap przygotowania modelu analitycznego z wykorzystaniem narzędzi wsparcia zakłada selekcję danych z wielu tabel źródłowego modelu relacyjnej hurtowni danych oraz ich transformację do postaci jednowierszowej.

4.4. Ontologie i metadane

Repozytorium ontologii i metadanych jest warstwą semantyczną rozwiązania, tj. przechowuje odwzorowanie pojęć dziedzinowych na identyfikatory obiektów bazy danych w bazie relacyjnej i NoSQL. Na podstawie danych zgromadzonych w tym repozytorium system pozwala zbudować zapytania sięgające do danych przy zastosowaniu pojęć „biznesowych”[5].

4.5. Narzędzia eksploracji danych

Narzędzia eksploracji danych umożliwiają przeprowadzanie analiz predykcyjnych (ang. *predictive analytics*), wyszukiwanie skupień, analizę asocjacji i sekwencji. Wspólną cechą tych narzędzi jest interfejs graficzny, umożliwiający definiowanie transformacji bez konieczności znajomości modelu danych oraz języków programowania. Narzędzia te umożliwią użytkownikom realizację następujących zadań:

- Określenie populacji, selekcję próby.
- Zbudowanie zapytań pobierających dane z hurtowni przypadków medycznych oraz transformujących dane do postaci pojedynczego rekordu (postać wymagana przez dostępne narzędzia eksploracji danych).
- Transformację danych – uzupełnienie brakujących wartości, detekcję i korektę wartości odstających, wprowadzenie efektów pochodnych.
- Wyodrębnienie ciągów: uczącego, testowego, walidacyjnego, próby kontrolnej.
- Indukcję reguły klasyfikacyjnej.
- Diagnostykę modelu – wartości miar oceny jakości predykcji, np.: wzmocnienie (*lift*), współczynnik błędnej klasyfikacji, indeksu ROC i innych.
- Przeprowadzanie procesu grupowania i klasyfikacji dokumentów tekstowych.

Narzędzia eksploracji danych będą głównym środowiskiem analitycznym rozwiązania, udostępnionym szerokiej grupie odbiorców. Dane uczące muszą być przygotowane w jednej tabeli relacyjnej bazy danych – naturalnym środowiskiem działania tych narzędzi jest repozytorium relacyjne systemu. Zbudowane klasyfikatory mogą zostać udostępnione użytkownikom, w szczególności

[5] A. Khan, J. Doucette, C. Jin, L. Fu, R. Cohen, *An ontological approach to data mining for emergency medicine*, w: *Proceedings of the 40th Annual Meeting Northeast Decision Sciences Institute Conference*, red. A.E. Avery, Northeast Decision Sciences Institute, Montreal 2011, s. 578–594.

lekarzom prowadzącym leczenie do wsparcia diagnostyki i planowania leczenia oraz na potrzeby edukacyjne.

4.6. Środowisko analityczne *big data*

Ponieważ reprezentacja danych w postaci atrybut-wartość wykorzystywana w „klasycznych" narzędziach eksploracji danych nie umożliwia uwzględnienia w danych wejściowych wielorelacyjnych zależności, część algorytmów analitycznych może być definiowana bezpośrednio na podstawie danych źródłowych z bazy danych NoSQL. Algorytmy takie mogą się odwoływać do nieustrukturalizowanej zawartości danych, jednakich budowa wymaga programowania w językach przetwarzania danych (Python, R lub pochodne).

4.7. Repozytorium modeli analitycznych

Repozytorium modeli analitycznych jest biblioteką, w której jest opisany każdy zbudowany model oraz w której są gromadzone dane o jego wykorzystaniu. Szczególnie istotnymi informacjami są miary jakości zbudowanych algorytmów, aktualizowane na podstawie zastosowania ich do nowych przypadków medycznych. Opublikowane w repozytorium modele analityczne tworzą katalog usług analitycznych świadczonych przez serwer usług analitycznych.

4.8. Serwer usług analitycznych

Serwer usług analitycznych jest komponentem umożliwiającym wykorzystanie zbudowanych modeli analitycznych przez szerszą grupę użytkowników na potrzeby oceny wskazanych przez nich przypadków medycznych. Udostępnione usługi mogą obejmować:

- Przeprowadzenie klasyfikacji przypadku medycznego.
- Wskazanie *k* najbardziej podobnych przypadków medycznych.
- Wizualizację przypadku medycznego w odniesieniu do wybranej populacji.
- Wskazanie najczęściej współwystępujących jednostek chorobowych.
- Wskazanie najczęściej stosowanych oraz najskuteczniejszych procedur medycznych[6].
- Wskazanie najczęściej występujących efektów ubocznych.

[6] G. Bliźniuk, T. Gzik, J. Koszela, *Dynamiczne ścieżki kliniczne*, „Biuletyn" WAT, nr 1, Wydawnictwo WAT, Warszawa 2013, s. 129–141.

Usługi będą mogły być wywoływane w dwóch trybach wprowadzania danych wejściowych:

- Wprowadzenie wszystkich cech wejściowych (atrybutów przypadku medycznego) z wykorzystaniem dedykowanego interfejsu (zakres cech jest określony przez wykorzystywany model) i uruchomienie procedury klasyfikacyjnej dla tych danych.
- Uruchomienie na podstawie danych już zapisanych w hurtowni danych – użytkownik wprowadzałby jedynie identyfikator przypadku medycznego. W tym trybie działania zadaniem usługi byłoby wyodrębnienie z hurtowni wszystkich danych wymaganych przez model.

Uruchomienie modelu analitycznego dla wskazanego przez użytkownika przypadku medycznego oznacza:

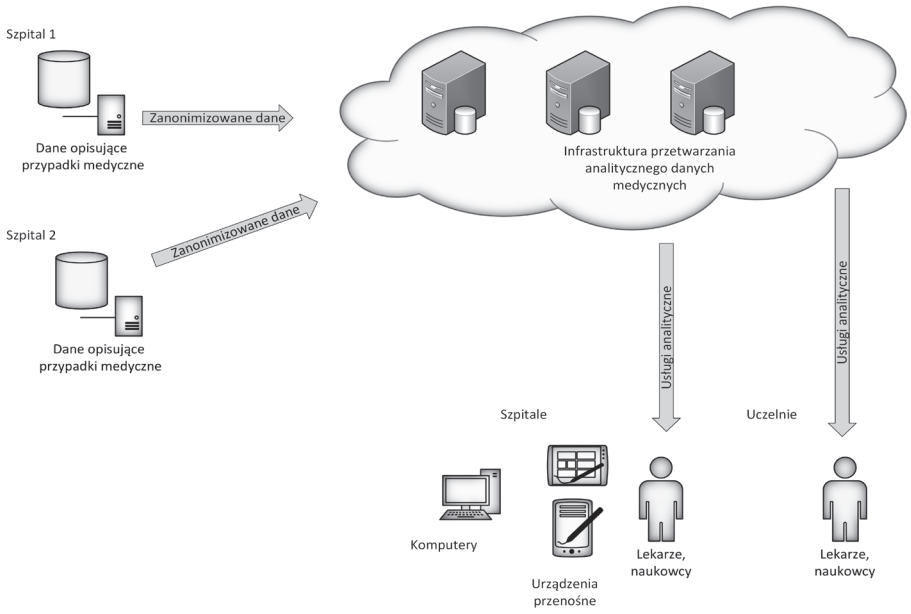
- Pobranie z repozytorium wszystkich danych wymaganych jako dane wejściowe związanych z przypadkiem medycznym. Specyfikacja danych wejściowych jest integralną częścią każdego z modeli, natomiast ich lokalizacja w repozytoriach jest określona w repozytorium ontologii i metadanych.
- Zbudowanie struktur tymczasowych, w których będą przechowywane pobrane dane wejściowe.
- Uruchomienie kodu procedury obliczeniowej.
- Aktualizację metadanych.
- Wizualizację wyników i udostępnienie ich użytkownikom.

5. Model usług analitycznych

Skuteczność systemów wspomagających podejmowanie decyzji zależy od jakości wprowadzonych danych oraz liczby przypadków, które stanowią podstawę do wnioskowania. Z drugiej strony, wdrożenie systemu analitycznego o przedstawionej architekturze wiąże się z dużymi nakładami finansowymi, z których zwrot nie zawsze jest widoczny ze względu na system finansowania służby zdrowia.

Rozwiązaniem obydwu problemów jest system analityczny funkcjonujący w chmurze, udostępniający usługi analityczne w modelu *Analytics as a Service*. Podstawą jego funkcjonowania jest wspólne repozytorium zasilane przez dane gromadzone przez wszystkie zainteresowane współpracą jednostki medyczne. Użytkownicy (lekarze i analitycy) mają możliwość budowy modeli i ich udostępniania, przy czym weryfikacja ich przydatności odbywa się na podstawie znacznie

szerszej populacji przypadków medycznych. Koncepcja tej infrastruktury została przedstawiona na rysunku 3.



Rysunek 3. Koncepcja budowy systemu analitycznego w chmurze
Źródło: opracowanie własne.

Redukcja kosztów budowy własnego środowiska analitycznego oraz dostęp do szerszej bazy przypadków medycznych będą czynnikami motywującymi do wykorzystania infrastruktury przez różne instytucje. Bezpośrednimi odbiorcami rezultatów projektu mogą być:

- Akademie medyczne oraz inne uczelnie medyczne – szkoły wyższe wyspecjalizowane w działalności dydaktycznej w dziedzinach nauk medycznych.
- Kliniki medyczne oraz szpitale kliniczne – specjalistyczne szpitale bądź oddziały szpitalne wchodzące w skład wyższych uczelni lub instytutów naukowych, które oprócz zapewnienia opieki medycznej oraz udzielania świadczeń zdrowotnych prowadzą prace dydaktyczne oraz naukowo-badawcze.
- Instytuty oraz centra medyczne – jednostki, których główny profil działalności obejmuje prowadzenie prac rozwojowo-badawczych oraz badań klinicznych.

Dostęp do usług analitycznych może zostać skomercjalizowany i stanowić źródło finansowania dla podmiotów uczestniczących w przedsięwzięciu. Oprócz

typowych usług, opisanych wcześniej i realizowanych na poziomie pojedynczego przypadku medycznego, wraz z rozszerzeniem zakresu integrowanych działań stają się możliwe:

- Monitoring i bieżąca analiza zagrożeń chorobowych społeczeństwa (grup zawodowych, wiekowych itp.).
- Generowanie najlepszych wariantów sekwencji badań specjalistycznych i konsultacji.
- Budowa symulatorów diagnostycznych do celów badawczych i dydaktycznych[7].

System analityczny udostępniający swoje usługi w chmurze, wychodzący poza dane, których właścicielem jest pojedyncza jednostka medyczna, wymusza powstanie modelu biznesowego funkcjonowania przedsięwzięcia. Powinien on uwzględniać system opieki zdrowotnej funkcjonujący w kraju oraz udział podmiotów zarówno publicznych, jak i prywatnych świadczących usługi komercyjne.

6. Podsumowanie i uwagi końcowe

Włączenie do architektury systemu analitycznego technologii *big data* pozwala osiągnąć korzyści z wykorzystania danych, które nie mogły zostać przetworzone w środowisku relacyjnej bądź wymiarowej hurtowni danych ze względu na wolumen, format oraz szybkość napływu. Pozostawienie w architekturze rozwiązania relacyjnego jest podyktowane bogatszą funkcjonalnością narzędzi eksploracji danych, które są rozwijane od dwóch dekad. Stosowane rozwiązania *big data* są młodą technologią, intensywnie rozwijaną, zarówno w formule otwartego oprogramowania (*open-source*), jak i przez dostawców komercyjnych. Implementacje systemów zarządzania bazami danych istotnie różnią się. Brakuje standardów pozwalających na abstrahowanie od sposobu implementacji, tak jak ma to miejsce w przypadku języków SQL lub MDX. W opisanej architekturze obydwie technologie uzupełniają się, udostępniając każdemu użytkownikowi widok na dane, dostosowany do jego potrzeb.

Połączenie obydwu sposobów przetwarzania danych w jednym rozwiązaniu rodzi wiele tematów badawczych. Najważniejszym problemem związanym z budową współdzielonego repozytorium danych jest zagadnienie integracji

[7] T. Nowicki, G. Bliźniuk, M. Lignowska, *Badanie efektywności procedur medycznych zapisanych w postaci ścieżek klinicznych*, „Roczniki" KAE, z. 25, Oficyna Wydawnicza SGH, Warszawa 2012, s. 37–54.

semantycznej danych pochodzących z różnych systemów źródłowych oraz możliwości automatycznej rozbudowy bazy wiedzy na podstawie napływających danych. Automatyzacja działań w tym zakresie może być czynnikiem sukcesu wdrożenia przedsięwzięcia ze względu na liczne ograniczenia związane z pozyskiwaniem tej wiedzy od ekspertów.

Nie mniej ważnym zagadnieniem projektowym jest realizacja mechanizmów odpersonalizowania danych w przypadku udostępnienia usług analitycznych większej liczbie podmiotów. Zagadnienie to obejmuje zarówno kwestie związane z lokalizacją repozytoriów utrzymujących odwzorowania kluczy wykorzystywanych w systemie na rzeczywiste identyfikatory, jak i sposób maskowania danych w źródłowych formatach danych przesyłanych do repozytorium.

Kluczowym czynnikiem powodzenia przedstawionego projektu analitycznego jest jakość zgromadzonych danych, a ta z kolei jest pochodną ergonomii interfejsów użytkownika, służących do wprowadzania danych. Komunikacja lekarzy z systemem może zostać usprawniona przez zastosowanie urządzeń mobilnych, umożliwiających wprowadzanie danych w miejscu ich powstawania (obchód, wizyta ambulatoryjna, konsylium), a także wykorzystanie możliwości wprowadzania danych za pomocą poleceń głosowych, bez konieczności wypełnienia formularzy. Motywacją do wykorzystania systemu powinna być dostępność on-line wyników badań dotyczących wskazanego przypadku.

Na koniec rozważań należy zaznaczyć, że tak skonstruowany system wspomagania decyzji nie podejmuje decyzji za lekarza. Wykorzystujący dane system wspomagania decyzji uzupełnia wiedzę i doświadczenie lekarza o mechanizmy wnioskowania statystycznego opartego na znacznie szerszej próbie przypadków medycznych niż próba wynikająca z doświadczeń jednostkowych lekarza.

Bibliografia

1. Bliźniuk G., Gzik T., Koszela J., *Dynamiczne ścieżki kliniczne*, „Biuletyn" WAT, nr 1, Wydawnictwo WAT, Warszawa 2013.
2. Emam K.E., *Guide to the De-Identification of Personal Health Information*, CRC Press, Boca Raton 2010.
3. Górski T., *Architektura platformy integracyjnej dla elektronicznego obiegu recept*, „Roczniki" KAE, z. 25, Oficyna Wydawnicza SGH, Warszawa 2012.

4. Khan A., Doucette J., Jin C., Fu L., Cohen R., *An ontological approach to data mining for emergency medicine*, w: *Proceedings of the 40th Annual Meeting Northeast Decision Sciences Institute Conference*, red. A.E. Avery, Northeast Decision Sciences Institute, Montreal 2011.
5. Marz N., *Big Data Principles and best practices of scalable realtime data systems*, MEAP Edition, Manning Publications, Greenwich, CT, USA 2012.
6. Nowicki T., Bliźniuk G., Lignowska M., *Badanie efektywności procedur medycznych zapisanych w postaci ścieżek klinicznych*, „Roczniki" KAE, z. 25, Oficyna Wydawnicza SGH, Warszawa 2012.
7. Savage N., *Better Medicine through Machine Learning*, „Communications of the ACM" 2012, vol. 55, no. 1.

Źródła sieciowe

1. http://www-01.ibm.com/software/data/bigdata/industry-healthcare.html (data odczytu 21.11.2013).
2. http://www.ehealthinformation.ca/knowledgebase/category/6/0/10/De-identification-Practices (data odczytu 21.11.2013).
3. http://www.healthcarebusinessintelligenceforum.com (data odczytu 21.11.2013).
4. http://www.technologyreview.com/news/518916/a-hospital-takes-its-own-big-data-medicine (data odczytu 21.11.2013).

* * *

Architecture of clinical decision support system using the Big Data concept

Summary

Clinical decision support systems based on relational and multidimensional technology lack the ability to process all available data because of its volume and format. On the other hand, NoSQL repositories offer great flexibility and speed in terms of data processing but require programming skills. The proposed solution is to combine both technologies in a single analytical system. A dual view of the data gathered in the repository allows to use data-mining tools, while the Big Data technology delivers the necessary data. Predictive models are then published as services in the Analytics-as-a-Service model to be run by medical staff in everyday practice. The unique feature of the solution is a possibility to score a medical case immediately, as the relevant data is available in the source systems.

Keywords: predictive analytics, medical decision support system, NoSQL