

ARTYKUŁY POGLĄDOWE (REVIEW PAPERS)

Wielokierunkowe wykorzystanie Big Data

(Multidirectional use of Big Data)

S Kasza ^{1,A,D}, A Romaszewski ^{1,E}, Z Kopański ^{1,D,F}, W Uracz ^{2,B,E}, F Furmanik ^{2,C}, S Dyl ^{2,B}, J Tabak ^{2,B}

1. Wydziału Nauk o Zdrowiu Collegium Medicum Uniwersytet Jagielloński
2. Collegium Masoviense – Wyższa Szkoła Nauk o Zdrowiu

Abstract - The authors have attempted to define a range of data that can be described as Big Data. They pointed out that the modern growth of records is a consequence of the rapidly growing public administration, financial, commercial, health and other segments of the economy. Informatization of health care connected with the insurance market is rapidly contributing to the generation of huge amounts of often sensitive data. Generating such a large amount of data requires the skillful use of analytical tools to extract the most-needed and most valuable information that will help us understand our behaviors, processes and improve the quality of life. These processes and tasks can be used by Big Data. Next, the authors discussed in detail the use of Big Data in the economy. They characterized the processing of large databases using the Big Data potential.

Key words - Big Data data range, use in the economy .

Streszczenie - Autorzy podjęli próbę zdefiniowania dokładnego zakresu danych, który można określić mianem Big Data. Podkreślili, że to zadanie może jednak stwarzać pewne trudności. Zwrócili uwagę, że współczesny rozrost rekordów jest następstwem gwałtownie rozwijającej się usługi administracji publicznej, sektora finansowego, handlowego, zdrowotnego i innych segmentów gospodarki. Informatyzacja ochrony zdrowia połączonego z rynkiem ubezpieczeń, gwałtownie przyczynia się do generowania ogromnych ilości niejednokrotnie wrażliwych danych. Generowanie tak dużej ilości danych wymaga umiejętnego zastosowania narzędzi analitycznych, aby wydobyć najpotrzebniejsze i najbardziej wartościowych informacji, które pozwolą zrozumieć nasze zachowania, zachodzące procesy i ulepszyć jakość życia. Tym procesom i zadaniom może służyć Big Data. Następnie autorzy szczegółowiej omówili wykorzystanie analizy Big Data w gospodarce. Scharakteryzowali przetwarzanie wielkich baz danych z wykorzystaniem potencjału Big Data.

Słowa kluczowe - zakres danych Big Data, wykorzystanie w gospodarce.

Wkład poszczególnych autorów w powstanie pracy - A-Koncepcja i projekt badania, B-Gromadzenie i/lub zestawianie danych, C-Analiza i interpretacja danych, D-Napisanie artykułu, E-Krytyczne zrecenzowanie artykułu, F-Ostateczne zatwierdzenie artykułu

Adres do korespondencji - Prof. dr Zbigniew Kopański, Wydziału Nauk o Zdrowiu Collegium Medicum Uniwersytet Jagielloński, Kraków, ul. Piotra Michałowskiego 12, PL-31-126 Kraków, e-mail: zkopanski@o2.pl

Zaakceptowano do druku: 29.08.2018.

ZDEFINIOWANIE OBJĘTOŚCI DANYCH OBJEJMUJĄCYCH DEFINICJE ZAKRESU BIG DATA

Zdefiniowanie dokładnego zakresu danych, który można określić mianem Big Data, może stwarzać pewne trudności. Według przytoczonej na początku pracy definicji Big Data to: Zakres informacji, który jest zbyt obszerny i skomplikowany do przetwarzania tradycyjnymi modelami. Cechą pojęcia według Gartnera jest duża objętość (Volume) danych, zatem spektrum będzie bardzo szerokie, najczęściej od

tysięcy do setek milionów pozycji, choć nie istnieje maksymalny analizowany obszar. W ubiegłych latach gwałtownie rosła liczba gromadzonych informacji. Wprowadzenie na rynek nowych możliwości do przetwarzania danych takich jak rozmaite sensory, urządzenia, samodzielne lub zdalne aplikacje przyczyniły się do znacznego nagromadzenia danych. Współcześnie ilość nagromadzonych i otrzymywanych informacji przerasta nasze możliwości analityczne. Na przykładzie przeglądarki Google i zindeksowanych stron wyraźnie widać, że sama

liczba stron tworzona w Internecie rośnie bardzo szybko. W 1998 było około 1 miliona stron internetowych. Już 2 lata później wartość ta sięgnęła 1 miliarda, natomiast w 2008 było to 1 trilion zindeksowanych stron [1,2]. Tak wielki wzrost jest indukowany w dużej mierze przez zyskujące wysoką popularność serwisy społecznościowe takie jak Facebook, Twitter, LinkedIn itp. Co więcej, znaczną ilość danych kreuja także telefony komórkowe, które w erze tzw. smartphonów, stały się kieszonkowym komputerem, który podłączony ciągle do globalnej sieci internetowej, wysyła znaczne ilości danych. Przewiduje się także, że tzw. Internet Rzeczy będzie powodował wzrost wysyłanych i gromadzonych danych do nieprzewidywanych wcześniej rozmiarów. Nowa infrastruktura i oprogramowanie będzie tworzone również dla tej całkowicie nowej gałęzi gospodarki. W tym wypadku z myślą o komunikacji człowiek-maszyna oraz maszyna-maszyna. Kreowanie interakcji pomiędzy urządzeniami stanie się kluczowym modelem biznesowym wielu podmiotów na rynku, a jeżeli na chwilę obecną szacuje się go na 70 mld urządzeń z czego jedynie 1% jest obecnie podłączone do sieci (np. bankomaty, kamery CCTV), to daje to szerokie pole dla twórców i analityków wytworzonych w ten sposób danych. Jedynie w 2015 roku wyprodukowano więcej danych w dziedzinie medycyny molekularnej, a niżeli w okresie od 1990 do 2005 roku [3]. Współczesna nauka nie jest w stanie nadążyć za przetwarzaniem szybko rosnących baz danych. Przyczyną tego jest brak odpowiednich algorytmów, a w kontekście przetwarzania informacji medycznych, brak realnej współpracy pomiędzy medycyną, biologią, a naukami informatycznymi [3-5]. Rozrost rekordów to także gwałtownie rozwijające się usługi administracji publicznej, sektora finansowego, handlowego, zdrowotnego i innych segmentów gospodarki. Informatyzacja ochrony zdrowia połączonego z rynkiem ubezpieczeń, gwałtownie przyczynia się do generowania ogromnych ilości niejednokrotnie wrażliwych danych. Generowanie tak sporej ilości danych będzie wymagało umiejętnego zastosowania narzędzi analitycznych, aby wydobyć najpotrzebniejsze i najbardziej wartościowych informacji, które pozwolą zrozumieć nasze zachowania, zachodzące procesy i ulepszyć jakość życia.[5,6]

WYKORZYSTANIE BIG DATA W GOSPODARCE

Istnieje wiele metod wykorzystania analizy Big Data w gospodarce. Praktycznie nie występuje część rynku, w którym analiza dużych danych nie miałaby zastosowania. Segment finansowy, detaliczny czy telekomunikacyjny to tylko część z długiej listy potencjalnych beneficjentów tego zagadnienia. Zakres najważniejszych przykładów zastosowań został wymieniony poniżej:

Sektor finansowy: Przykładem działań analitycznych w tym zakresie może być odkrywanie wiedzy i zależności przez banki oraz instytucje finansowe. Przeważnie wspomniane instytucje posiadają duże bazy danych o wysokiej jakości, które mogą z powodzeniem wykorzystać do przewidywania spłaty zobowiązań i tworzenia akcji kredytowych. Z ogromnej możliwości działań można wyróżnić także klasyfikacje i grupowanie klientów na potrzeby marketingu i itp. [1,4-7]

Sektor detaliczny: W przestrzeni sprzedaży detalicznej wdrożenie systemu eksploracji danych pomaga analizować informacje gromadzone przez programy sprzedażowe. Wykorzystywanie historii operacji, dane o transporcie towarów, konsumpcji i zrealizowanych usługach pomagają lepiej określać zachowania klientów. Kluczowym aspektem jest też optymalizacja biznesu i jego działań. Data mining posiada szerokie zastosowanie szczególnie w sklepach internetowych, gdzie najróżniejsze statystyki i dane są kompletne, wiarygodne oraz wysokiej jakości. Potentaci internetowi tacy jak Amazon, Allegro czy Ebay wykorzystują gromadzone dane, aby rekomendować produkty, które najbardziej będą odpowiadały jego oczekiwaniom w danej chwili. Niezwykle ciekawym przykładem opracowanym na polskim rynku jest przewidywanie trendów konsumpcyjnych na portalu allegro przez firmę TradeWatch, która potrafi wykazać wiele prawidłowości uzyskiwanych na stronie aukcyjnej [5-8].

Sektor ochrony zdrowia i ubezpieczeń: W dobie coraz bardziej zacieklej konkurencji na rynku ubezpieczeń, wygrywa instytucja, która potrafi w najbardziej wnikliwie przeanalizować ogrom wpływających danych na temat klienta, konkurencji oraz rynku. Przetwarzane dane mogą pomóc uzyskać przewagę konkurencyjną poprzez przewidywanie ryzyka operacyjnego, potencjalnego zakupu nowych polis przez klientów bądź wykrywanie niebezpiecznych i niosących ryzyko wzorów zachowań. Pod kątem zdrowia jest to min. eksploracja indywidualnego DNA,

aby poprawić i monitorować zdrowie i wiele innych rozwiązań, które zostaną przedstawione w dalszych rozważaniach. Możliwości wykorzystywania nowoczesnych technologii gromadzenia i przetwarzania danych jest wiele. Nie sposób wymienić wszystkich zastosowań nowej koncepcji do zdobywania wiedzy bez pominięcia czegośkolwiek. Oprócz wymienionych prężnie rozwijających się sektorów gospodarczych należy nadmienić też dynamicznie rosnący sektor telekomunikacyjny, obronny, wspomniany już wcześniej marketingowy, technologiczny, a także koncepcji Smart Cities, która skupia miasta skoncentrowane na zrównoważonym rozwoju i wykorzystywania nowoczesnych technologii do bardziej komfortowego i lepszego życia w metropoliach [6-8].

PROCES PRZETWARZANIA WIELKICH BAZ DANYCH

Tradycyjne metody do procesu przetwarzania danych często wykorzystują nowe technologie z nowoczesnym oprogramowaniem. Nie inaczej będzie także w przypadku Big Data. Jednak w sytuacji konfrontowania wielkiej liczby danych nie można mówić o prostym procesie przeprowadzanym na komputerze klasy PC z pojedynczym narzędziem do analizy danych. Główną różnicą pomiędzy zwykłym przetwarzaniem informacji, a analizą Big Data jest proces w jaki sposób się to odbywa. Z definicji wielkie bazy danych są bardzo duże i proces przerabiania danych musi zostać podzielony na kilka etapów. Kolejną różnicą jest częste korzystanie z centralnych repozytoriów, aby uzyskać szerszy wgląd na badane korelacje. Dostępne w chmurze platformy typu open source (wolne, bezpłatne oprogramowanie) takie jak Hadoop czy MapReduce pozwalają w pełni wykorzystać potencjał Big Data. [1,3,5]

Proces wydobywania wiedzy z rozległych zbiorów danych można podzielić na kilka etapów [19], [20].

Wyróżniamy [4-8]:

- Selekcja (ang. selection) – Selekcja określonych danych nawiązujące do wybranych kryteriów.

Przykład: Posiadane telefony komórkowe przez osoby, które można w ten sposób zdefiniować

- Wstępną obróbkę (ang. preprocessing) – Etap ten polega na początkowej analizie surowych danych pozyskanych z selekcji. Odrzuceniu wszystkich danych, które są zbędne i niepotrzebne w dalszym etapie.

Przykład: Kiedy badana jest zachorowalność na raka prostaty, nie trzeba zaznaczać płci

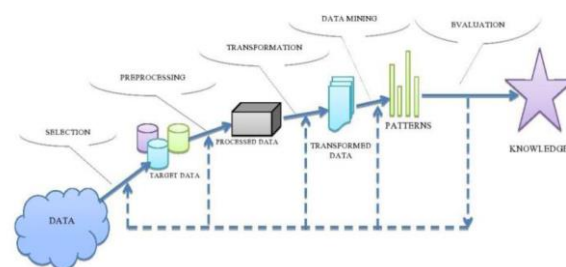
- Transformacje (ang. transformation) – Transformacja konkretnych danych, które odpowiadają kryterium postawionej hipotezy. Określa się tylko taki zakres danych jaki jest nam potrzebny do badania.

Przykład: Dopieramy wiek badanych użyteczny do przebadania konkretnego rynku

- Eksploracja danych (ang. data mining) – Na tym etapie następuje proces odkrywania wiedzy, czyli tworzenie konkretnych wzorów z pozyskanych danych

Przykład: Wykorzystanie algorytmu sztucznych sieci neuronowych do określenia wzorów

Interpretacje i ewaluacje (ang. interpretation and evaluation) – Wszystkie znaczące wzorce, które zostały wyodrębnione w poprzednim punkcie zostają zinterpretowane. Wiedza zdobyta w ten sposób może być wykorzystana do podjęcia konkretnych decyzji.



Rycina 1. Etapy procesu odkrywania wiedzy [4-8]

Omówione powyżej etapy procesu wymagają odpowiednich instrumentów oraz źródeł w celu ich realizacji. Można wyodrębnić cztery główne grupy narzędzi, które stanowią bazę do realizacji powyższych procesów. Nie da się jednoznacznie określić granic poszczególnych etapów, bowiem działania mogą zachodzić równolegle.[4,5] Jednak przybliżony schemat został podany poniżej.

I. Etap Selekcji – wyodrębnienie źródeł danych

Istotnym zagadnieniem w procesie przetwarzania dużych baz danych jest ich pozyskiwanie. Na tle obecnego tematu wyznaczenie źródeł informacji, zostanie ograniczone do ochrony zdrowia. Źródła surowych danych można podzielić na kilka rodzajów.

Pierwszą kategorią są wewnętrzne bazy danych np. elektroniczne rekordy zdrowotne (EHR) lub kliniczne systemy wsparcia np. CPOE {1}.

Drugim rodzajem są źródła zewnętrzne np. informacje rządowe, laboratoria, apteki, instytucje ubezpieczeniowe itp. {2}

W ramach kolejnej kategorii wyróżniamy różne formaty np. kartotekowe bazy danych, .csv, ASCII/tekst itd, bądź różne lokalizacje zarówno geograficzne jak i różne strony internetowe świadczeniodawców. {3}

Odrębnym i zarazem ostatnim wariantem są różne aplikacje np. bazy danych, aplikacje przetwarzające transakcje itp. {4}

Do źródeł danych zalicza się także wszelkie strony internetowe i portale społecznościowe oraz aplikacje mobilne. {5}

uwagi na rosnącą popularność wymienionych źródeł cyfrowych przewiduje się, że w przyszłości będą one zyskiwać na znaczeniu. Jednak przeszkodą w ich stosowaniu na użytek publiczny i do badań naukowych jak i komercyjny, może być stan właścicielski i fakt, że są one najczęściej usługami prywatnymi. W przyszłości dużą rolę będzie odgrywał też wspomniany wcześniej Internet Rzeczy. {6} Umożliwi on gromadzenie danych maszynowych, odczytów z urządzeń sensorycznych i systemów mierzących funkcje życiowe. (tabela 1)

II. Etap wstępnej obróbki i Transformacji

W ramach dalszych etapów obróbki danych należy wspomnieć również o oprogramowaniu pośredniczącym (ang. middleware), które jest częścią procesu pozyskiwania i przetwarzania danych. W tym wypadku będą to tzw. ETL, (ang. extract transform load), hurtownie danych, tradycyjne formaty typu CSV oraz tabele. [4-8]

III. Etap Wstępnej obróbki i Transformacji

Nieodłącznym elementem na drodze do poznania nowej wiedzy, będą też platformy i narzędzia do Big Data takie jak: Hadoop, MapReduce, Pig, Hive, Jaql, Zookeeper, HBase, Cassandra, Oozie, Avro, Mahout oraz wiele innych rozwiązań.[4-9]

IV. Etap Ekploracji danych oraz Interpretacji i Ewaluacji

Ostatecznym etapem, który wymaga odpowiednich technik i oprogramowania jest proces przekształcania danych w informacje, a informacje w wiedzę. Najczęściej zdefiniowany etap jest określany mianem Business Intelligence (BI). To moment, w

którym wykorzystuje się szereg narzędzi analitycznych takich jak OLAP czy eksploracja danych do pozyskania wiedzy, a także raporty, prezentacje, czy wizualizacje obrazkowe do przekazania informacji. Popularnym narzędziami są również gotowe oprogramowania, zawierające wszystkie niezbędne procesy i narzędzia w ramach dostarczonej usługi. W tej chwili najczęściej używane to Teradata czy IBM DB2, natomiast popularność zyskują rozwiązania chmurowe. [1,4-9]

Tabela 1. Grupy źródeł i narzędzi kluczowych do analizy Big Data Opracowanie własne na podstawie [4,5,8,9]

I. Źródła surowych danych dla ochrony zdrowia	II. Oprogramowa nie pośredniczące	III. Narzędzia i platformy do Big Data	IV. Oprogramowa nie analityczne i prezentacja
{1} Wewnętrzne: Elektroniczne rekordy zdrowotne; Kliniczne systemy wsparcia; {2} Zewnętrzne: informacje rządowe, laboratoriów, apteki, instytucji ubezpieczeń; {3} Różne formaty: kartotekowe bazy danych; Różne lokalizacje: geograficzne, strony internetowe świadczeniodawc ów {4} Różne aplikacje: aplikacje przetwarzające transakcje, bazy danych {5} Strony internetowe; Portale społecznościowe; Aplikacje mobile; {6} Internet Rzeczy;	ETL; Hurtownie danych; CSV; Tabele;	Hadoop; MapReduc e; Pig; Hive; Jaql; Zookeeper ; HBase; Cassandra ; Oozie; Avro; Mahout;	Analityka: Teradata; IBM DB2; IBM SPSS; Rozwiązania chmurowe; Prezentacja: Raporty; Prezentacje; Wizualizacje;

PIŚMIENICTWO

1. Fan W, Bifet A. Mining Big Data: Current Status, and Forecast to the Future. New York; Wyd. ACM SIGKDD Explorations Newsletter 2012.
2. Kamal N, Wiebe S, Engbers J, Hill M D. Big Data and Visual Analytics in Health and Medicine: From Pipe Dream to Reality. Health Med Inform 2014; 5,5: 25-34.
3. Papińska-Kacperek J. Usługi cyfrowe. Perspektywy wdrożenia i akceptacji cyfrowych usług administracji publicznej w Polsce. Łódź; Wyd. Uniwersytetu Łódzkiego, 2013.
4. Diebold F. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive. Pennsylvania; Penn Institute for Economic Research, Department of Economics. University of Pennsylvania, 2012.
5. Marconi K, Dobra M, Thompson C. The use of Big Data in Healthcare. In: Liebowitz J. Big Data and Business Analytics. Boca Raton; CRC Press, 2013: 229–248.
6. Schmarzo B. Big Data. Understandinga How Data Powers Big Business. Indianapolis; John Wiley & Sons Inc., 2013.
7. Chen H, Chiang R H, Storey V C. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly 2012; 36, 4:1165–1188.
8. Ahmad P, Qamar S, Syed Qasim AR. Techniques of Data Mining in Healthcare: A Review. Int J Comput Applic 2015;120,15, 16-24.
9. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inform Sci Syst 2014; 2: 3-10.