JCSI 1 (2016) 6-13
Journal of Computer Sciences Institute

Wysłane: 2016-09-14
Przyjęte: 2016-09-18

Integracja Big Data i Business Intelligence jako innowacyjne rozwiązanie wspomagające funkcjonowanie nowoczesnych organizacji

Łukasz Bielak*, Piotr Muryjas*
Politechnika Lubelska, Instytut Informatyki, Nadbystrzycka 36B, 20-618 Lublin, Polska

Streszczenie. Celem niniejszego artykułu jest przedstawienie możliwości integracji Business Intelligence(BI) i Big Data (BD). Na podstawie studiów literaturowych określono potencjalne korzyści i wady płynące z takiego rozwiązania. Ponadto, wymienione zostały korzyści jakie odniosły organizacje, które wdrożyły rozwiązania BI i BD. Praca obejmuje także autorski projekt systemu integrującego BI i BD w organizacji z sektora medycznego.

Słowa kluczowe: Big Data; Business Intelligence; Big Data Analytics

*Autor do korespondencji

Adresy E-mail: lukasz.bielak1@gmail.com, p.muryjas@pollub.pl

Integration of Big Data and Business Intelligence as innovative solution supporting the functioning of modern organizations

Łukasz Bielak*, Piotr Muryjas*
Institute of Computer Science, Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland

Abstract. The purpose of this article is to present the possibilities of integration of Business Intelligence (BI) and Big Data (BD). Based on literature studies identified the potential benefits and drawbacks coming from this solution. In addition the benefits of this solution were listed, based on case studies. The article also includes a proprietary system design that integrates BI and BD in the organization of the medical sector.

Keywords: Big Data; Business Intelligence; Big Data Analytics

*Corresponding author

E-mail addresses: lukasz.bielak1@gmail.com, p.muryjas@pollub.pl

1. Wstęp

Obecnie niemal każda organizacja musi dbać o swoją innowacyjność, aby być konkurencyjną na rynku. Nowoczesne oprogramowanie, takie jak systemy CRM czy ERP ułatwiają prowadzenie i zarządzanie nowoczesnych firm. Takie systemy gromadzą dane na temat sprzedaży produktów, przeprowadzonych transakcji, czy zachowań klienta. W sytuacji, gdy organizacja posiada zarchiwizowane dane z jakiegoś okresu, korzystna może okazać się ich analiza. Do tego celu wykorzystuje się rozwiązania klasy Business Intelligence (BI), które na podstawie danych wspomagają podejmowanie decyzji biznesowych, wykrywają potencjalne przekłamania i próby oszustw. Warunkiem poprawnego działania systemu BI są wiarygodne dane, pochodzące najczęściej z systemów transakcyjnych danej firmy. Bardzo powszechny w dzisiejszych czasach dostęp do urządzeń mobilnych opartych o system operacyjny jest przyczyną generowania dużej ilości danych. Co za tym idzie, mogą one okazać się użyteczne w kontekście analizy i wydobywania wiedzy. Wadą takiego rodzaju danych jest często brak określonej struktury i szybkość ich generowania. W tradycyjnym systemie BI może okazać się niemożliwe przeanalizowanie tak wielkiej ilości danych. Zjawisko przyrostu dużej ilości różnorodnych danych w szybkim tempie nazywane jest obecnie Big Data (BD). Potencjalnie korzystne dla organizacji posiadającej system klasy BI może okazać się połączenie go z Big Data. Celem niniejszej pracy jest przedstawienie metod integracji Business Intelligence i Big Data, a także płynących z tego korzyści w oparciu o badania literaturowe i studia przypadków. Przytoczone też zostaną przypadki wdrożenia Business Intelligence i Big Data oraz ich wpływ na funkcjonowanie nowoczesnych organizacji. Ponadto, praca zawiera także autorski projekt systemu integrującego Business Intelligence i Big Data w organizacji z sektora medycznego.

2. Business Intelligence

Definicja zamieszczona w *IT Glossary* [20] określa terminem Business Intelligence (BI, system klasy Business Intelligence) aplikacje, infrastrukturę oraz najlepsze praktyki umożliwiające dostęp do danych i ich analizę. Celem omawianego rozwiązania jest optymalizacja i poprawa wydajności procesów biznesowych. Wang C. H. w artykule *A novel approach to conduct the importance-satisfaction analysis for acquiring typical user groups in business-intelligence systems* [13] podkreśla, że system BI ma za zadanie wspomaganie podejmowania decyzji. Podstawowym elementem architektury rozwiązań BI jest hurtownia danych(HD). W książce *Exam 70-463:Implementing a Data Warehouse with Microsoft® SQL Server® 2012* [11] jej autorzy Sarka D., Lah M. i Jerkic G. definiują hurtownie danych jako zcentralizowaną składnicę danych, która przechowuje scalone(z różnych systemów), oczyszczone, historyczne dane. Struktura HD ma zapewnić dużą

wydajność odczytu rekordów, przez co wykorzystuje się głównie trzy schematy logiczne przy jej projektowaniu, gwiazdy, płatka śniegu i konstelacji. Ładowanie danych do HD odbywa się cyklicznie i jest realizowane przez proces ETL, w którego skład wchodzą takie elementy jak ekstrakcja (ang. *extract*), transformacja (ang. *transform*) i ładowanie (ang. *load*) danych. Relatywnie, najbardziej skomplikowana jest transformacja, w której skład wchodzi zmiana struktury danych, a także zagregowanie określonych wartości transakcyjnych. R. Kimball w książce *The Data Warehouse Toolkit Third Edition* [5] zwraca uwagę, że budowa procesu ETL pochłania najwięcej czasu przy budowaniu systemu Business Intelligence. Elementem BI są także bazy analityczne OLAP (ang. *Online Analytical Processing*), które są zoptymalizowane pod kątem kwerend i raportowania, przez co możliwe są do wykonania operacje, które ze względów wydajnościowych nie byłyby możliwe w tradycyjnej hurtowni danych. W literaturze pojawiają się propozycje implementacji OLAP w chmurze obliczeniowej [14]. Kolejnym opisywanym w tej pracy elementem BI jest drążenie danych (ang. *Data Minig*). Proces ten, zgodnie z definicją proponowaną przez firmę ORACLE [26] polega na odkrywaniu zależności niewidocznych dla badacza, na podstawie zbiorów danych. W literaturze pojawiają się zastosowania Data Miningu w medycynie [1] i systemach ostrzegania o katastrofach naturalnych [7]. Najwyższą warstwą w systemach klasy BI jest warstwa prezentacji danych. Obecnie coraz częściej do tego celu wykorzystywane są narzędzia typu Self Service Business Intelligence, których działanie ma być jak najbardziej intuicyjne dla użytkownika. Taka filozofia pozwala na korzystanie z zawartości baz danych osobom, które nie posiadają specjalistycznej wiedzy. Dzięki aplikacją Self Service BI, można się w relatywnie prosty sposób podłączyć do źródła danych i przy pomocy metody przeciągnij i upuść zbudować wykres czy raport.

2.1. Kierunek rozwoju BI

Jak pokazuje raport wydany przez organizację TDWI, *TDWI Best Practises Report - Business Driven Business Intelligence and Analytics. Achieving Value through Collaborative Business/IT Leadership* [12] obecnie, najbardziej wymagającym elementem, z punktu widzenie utrzymania BI jest warstwa technologiczna. W jej skład wchodzą między innymi procesy ETL i hurtownia danych, a utrzymaniem tych elementów zajmują się osoby ze specjalistyczną wiedzą techniczną. Zupełnie inaczej wygląda obsługa warstwy prezentacji, z którą bez problemu radzą sobie analitycy i pozostałe osoby pracujące z narzędziami tego typu. Co za tym idzie widoczny jest trend rozwoju intuicyjnych narzędzi dla osób nieposiadających specjalistycznej wiedzy, który Firma *Tableau,* producent oprogramowania klasy BI, wymienia w publikacji *Top 10 business intelligence trends for 2016*[23].

2.2. Korzyści z wdrożenia systemu klasy BI

Wiele organizacji wdraża u siebie rozwiązania klasy BI, co pokazują Hannula M. i Pirttimaki V. w publikacji *Business intelligence empirical study on the top 50 Finnish companies* [4]. Artykuł opierał się na ankietach wypełnionych przez 50 fińskich firm mających najlepsze wyniki finansowe w 2002 roku. Głównymi powodami, dla których organizacje decydowały się na wdrożenie systemu BI były:

- Wspieranie podejmowania decyzji i planowanie;
- Wiedza o własnym biznesie i podejmowanie akcji;
- Bycie konkurencyjnym na rynku pracy;
- BI jest trendem na rynku biznesowym;

Znaczna większość organizacji uznała, że wdrożenie systemu BI przyniosło korzyści, co więcej 1/3 respondentów udzieliła odpowiedzi, że nakłady na tego typu systemy wzrosną znacznie. Natomiast 44% uważa, że wzrosną trochę. Korzyści z wdrożenia rozwiązania klasy BI odniósł także Santander Consumer Bank, studium przypadku zostało opisane w publikacji *Studium przypadku z wdrożenia systemu do budżetowania InForum BI Studio w Santander Consumer Bank* [22]. Za główne zalety systemu są uważane:

- Efektywność, proces tworzenia budżetu angażuje mniejsze zasoby ludzkie, sprzętowe i licencyjne;
- Elastyczność i skalowalność;
- Automatyzacja pracy, wydajność pracy, kontrola kompletności, spójności i jakości danych źródłowych;
- Szybkość i łatwość pracy, raporty są budowane w czasie wielokrotnie krótszym niż poprzednio;
- Samodzielność pracy analityków - nie jest potrzebna wsparcie ze strony IT;
- Intuicyjny interfejs;

3. Big Data

Intensywny rozwój techniki w ostatnich latach sprawia, że na rynku pojawia się coraz więcej urządzeń mobilnych bazujących na systemie operacyjnym. Takie rozwiązanie, w połączeniu licznymi czujnikami obecnymi w telefonach, czy tabletach wpływa na wzrost liczby aplikacji, które w założeniach mają ułatwić życie użytkownikowi. Konsekwencją tego zjawiska jest generowanie bardzo dużych ilości danych, stanowiących doskonałe podłoże do wydobywania z nich informacji, a ostatecznie wiedzy. Na uwagę zasługuje fakt, iż w omawianym przypadku, klasyczny model przechowywania danych oparty o bazę relacyjną nie sprawdzi się i koniecznie staje się inne podejście do tego zagadnienia. Nowoczesne składnice danych powinny obsługiwać dane nieustrukturyzowane, pochodzące z różnych źródeł, a także oferować szybkie rezultaty zapytań i możliwość analizy, takie rozwiązanie określane jest jako Big Data(BD). Mach-Król M. w publikacji *ANALIZA I STRATEGIA BIG DATA W ORGANIZACJACH* [6], zauważa, że obecnie nie ma ściśle określonej definicji terminu Big Data, zjawisko to jest opisywane za pomocą charakterystyk. Do opisu Big Data według Mach-Król M. wykorzystane są 3 główne właściwości:

- **Volume** (pol. objętość) cecha związana ze stale rosnącą ilością danych, która jest znacznie większa niż w klasycznych rozwiązaniach opartych o relacyjne bazy danych.
- **Velocity** (pol. szybkość) poprzez sieć internet stale przesyłane są dane, a przepustowość łącz musi być ciągle zwiększana. Wynika to z faktu wielu aplikacji mobilnych,

które w relatywnie krótkim czasie generują dane i transferują je do serwerów, bądź innych użytkowników.
- **Variety** (pol. różnorodność) właściwość, która określa strukturę danych. W Big Data gromadzone dane są w postaciach ustrukturyzowanych(systemy CRM, ERP itd.), semi-ustrukturyzowanych (EMail, facebook, twitter) i nieustrukturyzowanych(obrazy, dźwięki, filmy).

Wymieniona też zostaje 4 cecha **Veracity** (pol. wiarygodność), która pojawiła się niedawno i paradoksalnie wiąże się z niepewnością danych. Firma SAS Institute, zajmująca się wytwarzaniem oprogramowania z zakresu między innymi Business Intelligence, oprócz 3 wyżej wymienionych głównych cech BD wymienia w publikacji *Big Data - What it is and why it matters* [17] jeszcze dwie dodatkowe:

- **Variability** (pol. zmienność) ilość generowanych danych może rosnąć lub maleć w danym okresie, wpływ na to mają na przykład wydarzenia społeczne takie jak igrzyska olimpijskie czy finały konkursów artystycznych (duża liczba komentarzy i zdjęć na portalach społecznościowych).
- **Complexity** (ang. złożoność) dane pochodzą z różnych źródeł, a ich integracja i oczyszczanie jest skomplikowane i trudne.

3.1. Narzędzia i usługi do organizacji danych

Duże ilości danych, które w szybkim tempie zwiększają swoją objętość wymagają innego podejścia do ich przetwarzania. Według *Big Data for Dummies* [8], narzędzia i usługi do organizacji danych stanowią ekosystem technologii służących do integracji danych i przygotowania ich do dalszej analiz, w jego skład wchodzą:

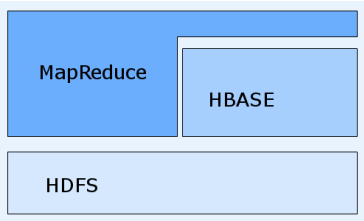
- rozproszony system plików zapewniający skalowalność i dużą przestrzeń dyskową,
- usługi serializacji niezbędne do zdalnego wywoływania procedur w systemie,
- usługi do koordynacji procesów rozproszonych,
- narzędzia ETL,
- usługi do zarządzania harmonogramem wykonywania procesów.

W *Big Data for Dummies* [8] autorzy zwracają uwagę na platformę Hadoop, jako przykład zbioru produktów do organizacji danych. Zgodnie z definicją podaną przez producenta w *Welcome to Apache Hadoop* [24] jest to framework, pozwalający na rozproszone przetwarzanie dużych zbiorów danych w klastrach komputerowych, z wykorzystaniem prostych modeli programowania. Zaprojektowany z myślą o skalowalności od jednego komputera do tysięcy maszyn. W tym rozwiązaniu, bardziej niż stosowanie elementów sprzętowych o wysokiej dostępności liczy się redundancja węzłów, aby w razie awarii, inna maszyna przejęła zadania uszkodzonej. Praca *Turning Small Analytics on Big Data: Data Partitioning and secondary indexes in Hadoop ecosystem* [10] prezentuje architekturę logiczną Hadoopa (rysunek 1), zawiera ona 3 główne elementy składowe: HDFS, HBase i MapReduce.

*Rozproszony system plików HDFS(*ang. *hadoop distributed)* - został zaprojektowany do przechowywania dużych plików na wielu maszynach. Ponadto, każdy plik jest podzielony na bloki, a te są replikowane na różne maszyny, przez co ryzyko awarii systemu spada.

HBase - jest to nierelacyjna, rozproszona, wersjonowana baza danych typu klucz-wartość. Umożliwia skalowanie horyzontalne (im więcej sprzętu, tym wydajniej działa).

MapReduce - framework programistyczny wykorzystywany przy tworzeniu rozwiązań obliczeń rozproszonych. Zasada działania opiera się na tworzeniu par klucz-wartość(*Map*), a następnie redukowaniu nieistotnych, z punktu widzenia zapytania wyników. Na rysunku 2 przedstawiony został przykład algorytmu MapReduce. Celem, jest uzyskanie zsumowanych wartości dla rekordów, których nazwy kontynentów są w zbiorze ('Europe', 'Africa'). Idąc od góry, dostępne są dwa zbiory danych zawierające klucz, nazwę kontynentu oraz wartość liczbową. Pierwszym krokiem, jest zmapowanie wartości, oznacza to, że do każdej nazwy kontynentu (zadeklarowanego wcześniej) przypisana zostaje liczba(Map). Zadanie to jest wykonywane na obu zbiorach równolegle. Kolejnym krokiem jest scalenie zbiorów(Merge-Sort). Przedostatnim etapem, jest zastosowanie funkcji redukcji(Reduce), która w tym wypadku, sumuje wartości, na przykład, wiersz <AFRICA,12,10 > po redukcji będzie miał postać <AFRICA,22 >. Ostatnim krokiem jest zwrócenie wyników, czyli w tym przypadku wierszy <AFRICA,22 > i <Europe, 15>



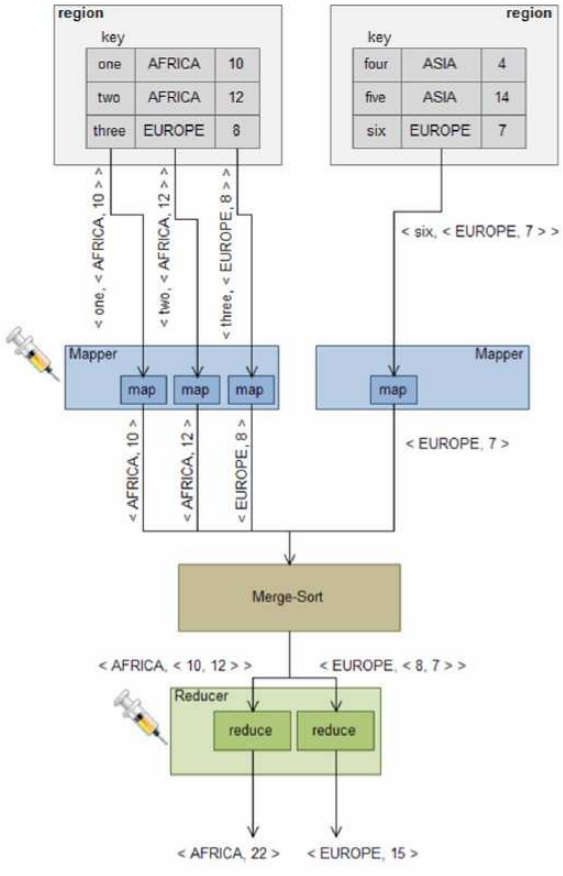
Rys. 1. Architektura logiczna Hadoopa na podstawie [10]

Dzięki wykorzystaniu opisywanych w tym rozdziale narzędzi, możliwa jest techniczna implementacja systemu obsługującego Big Data.

3.2. Big Data w organizacjach

Rosnąca popularność Big Data powoduje, że coraz więcej organizacji interesuje się rozwiązaniami tego typu. *International Institute for Analytics* opublikował raport *Big Data in Big Companies* [3] który został przygotowany na podstawie wywiadów przeprowadzonych w 2013 roku z dużymi organizacjami. Głównym celem było określenie, jak Big Data dostosowało się do już działających środowisk analitycznych, a także składujących dane. Jak zostało wspomniane wcześniej, 3 cechy określające Big Data to objętość, różnorodność i szybkość. Raport pokazuje, że najbardziej istotna dla firm jest różnorodność, która umożliwia analizę danych z różnych źródeł. Co może

potencjalnie przynieś zysk. Z uwagi na ciągły rozwój technologii, źródła danych takie jak strony internetowe czy logi są coraz bardziej powszechne i niosą ze sobą informację. Przykładem jest firma United Healthcare, która wykorzystuje Big Data do takich czynności jak wspieranie planowania leczenia. Co więcej, do kontroli jakości usług wykorzystywane są tam nagrania rozmów telefonicznych zmieniane na tekst i poddawane ostatecznie analizie z wykorzystaniem przetwarzania języka naturalnego. Jak przyznaje organizacja, kiedyś takie rozwiązanie nie byłoby możliwe, jednak obecne metody analizy pozwalają na określenie czy klient wyraża się pozytywnie, czy negatywnie o firmie. Dane typu Big Data i ich analiza we współczesnych organizacja przyczyniają się do redukcji czasu, jaki kiedyś był potrzeby do wykonania pewnych zadań optymalizacyjnych. Firma Macy's, która jest właścicielem sieci galerii handlowych w Stanach Zjednoczonych informuje, że proces optymalizacji cen dla 73 milionów produktów został skrócony z 27 do 1 godziny. Jest to rezultat implementacji klastra Hadoop i zastosowania obliczeń rozproszonych, dodatkową korzyścią, jest redukcja kosztów sprzętu o 70%, co wynika z możliwości skalowania horyzontalnego Hadoopa. Firma UPS, zajmująca się przewozem przesyłek i logistyką dane na temat transakcji i tras przewozów gromadzi od 1980 roku. Samochody, którymi poruszają się kurierzy posiadają urządzenia zbierające informacje o dziennej wydajności(szybkość, kierunek, hamowanie). Optymalizowanie tras na podstawie takich danych doprowadziło do zaoszczędzenia 8.4 milionów galonów paliwa w 2011 roku, co w konsekwencji przyczyniło się do redukcji kosztów. Planowane jest także wdrożenie takiego rozwiązania dla floty lotniczej. Oprócz redukcji koszów i czasu, powstają też całkiem nowe zastosowania wielkich zbiorów danych. Przykładem tego są oferty konstruowane na podstawie zgromadzonych danych (Data-Based offering). Firmy takie jak T-Mobile, Sprint, Verizon Wireless działające w branży komunikacyjnej, na podstawie zachowań użytkowników urządzeń mobilnych i ich lokalizacji przygotowują oferty dostosowane do regionów. Netflix jest jedną z największych firm oferujących wypożyczanie filmów przez media strumieniowe. System rekomendacji video jaki jest oferowany użytkownikom, również opiera się o Big Data. Powyższe rozwiązania są nie tylko korzyścią dla organizacji w postaci większych zysków ale także dla klientów, gdyż proponowana jest im oferta zbliżona do potrzeb, może się to przekładać na wzrost zadowolenia z usług. Podobnie jak w przypadku Busienss Intelligence, jednym z zadań Big Data jest wspomaganie podejmowania decyzji biznesowych, jednak w drugim przypadku wykorzystywane są także dane nieustrukturyzowane. Organizacje wykorzystujące do tego celu Big Data, korzystają z logów aplikacji webowych, rozmów telefonicznych, czy korespondencji mailowych. Amerykańskie banki Wells Fargo, Bank of America i Discover przyznały, że z uwagi na wiele kanałów, jakimi można dokonać zakupu usługi, monitorują postępowania swoich klientów. Dzięki temu możliwe jest lepsze zrozumienie grup społecznych decydujących sie na skorzystanie z usługi, a w konsekwencji optymalizacje procesów sprzedaży. Organizacje starają się zwiększać swoje zasoby ludzkie związane z zaawansowaną analizą danych (Data Science). Najchętniej poszukiwane są osoby z dobrymi umiejętnościami Data Science, manipulowania danymi typu Big Data i IT umożliwiającymi pisanie skryptów. Pracodawcy zwracają uwagę, że najbardziej istotną kwestią jest umiejętność prezentacji danych w formie graficznej, czy narracyjnej. Bardziej liczy się przekazanie wyników badań w sposób przystępny niż znajomość narzędzi. GE amerykański konglomerat, działający między innymi w branżach energetycznych i produkcji maszyn, dla osób nowozatrudnionych na stanowisko Data Scientist ma przygotowany wewnętrzny program szkoleń.



Rys. 2. Algorytm MapReduce, źródło : Turning Small Analytics on Big Data: Data Partitioning and secondary indexes in Hadoop ecosystem [10]

4. Big data Analytics i integracja Business Intelligence i Big Data

Firma SAS Institute definiuje Big Data Analytics (BDA) jako badanie wielkich zbiorów danych w celu odkrycia ukrytych wzorców, korelacji i innych zależności (*Big Data Analytics - What it is and why it matters*) [16]. Z kolei IBM w publikacji *What is Big Data Analytics* [25] określa BDA jako użycie zaawansowanych technik analitycznych na wielkich i różnorodnych zbiorach danych, włączając w to dane ustrukturyzowane i nieustrukturyzowane, a także pakiety i strumienie. Taka analiza pomaga badaczom i użytkownikom biznesowym na trafniejsze i szybsze podejmowanie decyzji z wykorzystaniem Big Data. Zaawansowanie techniki analizy takie jak analiza tekstu, uczenie maszynowe, analiza predykcyjna, data mining, statystyka i przetwarzanie języka naturalnego skutkują znacznie trafniejszym i szybszym

decyzjom biznesowym. Podobne techniki stosuje się także w przypadku systemów klasy Business Intelligence. Jak podają autorzy *Big Data for Dummies* [8] kluczowe różnice między dwoma systemami, BI i BDA to przede wszystkim :

- dane typu Big Data mogą pochodzić z niepewnych źródeł,
- dane typu Big Data mogą być nieoczyszczone,
- dane typu Big Data mogą mieć niski współczynnik sygnału do szumu,
- Big Data Analytics umożliwia analizę w czasie rzeczywistym.

J., Hurwitz, A., Nugent, F., Halper, i M., Kaufman w *Big Data for Dummies* [12] dzielą metody analizy danych na 4 podgrupy:

- Prosta analiza, która składa się z takich metod jak
  - wycinanie (ang. slicing and dicing) polegająca na zawężaniu zbiorów danych do interesujących użytkownika na przykład tylko dane z 2016 roku,
  - proste monitorowanie zmian w czasie rzeczywistym,
  - identyfikacja anomalii.
- Zaawansowana analiza, składająca się z:
  - modelowania predykcyjnego pomagającego w obliczaniu prawdopodobieństwa zdarzeń w przyszłości. Zastosowaniem może być wykrywanie, czy transakcja kartą kredytową może być próbą oszustwa,
  - analizy tekstu,
  - innych statystycznych algorytmów takich jak segmentacje i mikrosegmentację,
  - data miningu.
- Analiza operacyjna używana na przykład przy dobieraniu odpowiedniego produktu dla klienta.
- Monetyzacja danych (ang. *Monetizing analytics*) polega na zamianie zgromadzonych informacji na zysk w postaci pieniądza, wykorzystywane między innymi do odpowiedniego rozlokowania punktów sprzedaży na podstawie wiedzy o klientach.

4.1. Self Service BI jako narzędzie BDA

Podobnie jak w systemach klasy Business Intelligence, zgromadzone dane należy przedstawić w formie, która przenosi jak najwięcej treści dla użytkownika końcowego. Z tego względu, niektóre aplikacje typu Self Service BI umożliwią podłączenie się do Hadoopa, bądź baz typu *NoSQL*. Producenci narzędzi BI firmy *Tableau* i *Microsoft* zaimplementowali w swoich produktach możliwość podłączenia się do źródeł typu Big Data (*Connecting to Hadoop Hive, Data sources in Power BI Desktop*) [18, 19], a także z bazami *NoSql (Tableau & MongoDB: Visual Analytics on JSON at the Speed of Thought)* [23]. Dzięki temu, możliwe staje się zintegrowanie ze sobą danych typu Business Intelligece i Big Data. Konsekwencją tego jest relatywnie proste i intuicyjne tworzenie kokpitów menedżerskich, a także raportów przedstawiających dane z różnych źródeł. Na przykładzie tych dwóch wymienionych produktów zauważyć można, że aplikacje typu Self Service BI sprawdzają się także jako aplikacje Big Data Analytics.

4.2. Zaawansowana analiza z użyciem platform BDA dostępnych na rynku.

Analiza predykcyjna i inne działy zaawansowanej analizy są obecnie najprężniej rozwijającym się przedstawicielem rynku analityki dla firm według raportu Gartnera [30]. Zgodnie z definicją przedstawioną w omawianym dokumencie, zaawansowana analityka to analizowanie wszystkich typów danych z wykorzystaniem wysokiej klasy metod ilościowych (statystki opisowe i predykcyjne, data mining, uczenie maszynowe, symulacja i optymalizacja) dla uzyskania wiedzy, której nie udałoby się wydobyć z wykorzystaniem klasycznego BI (raporty i zapytania). Na rynku obecnych jest relatywnie dużo rozwiązań spełniających powyższe kryteria, co pokazuje rysunek 4.1 Autorzy raportu podkreślają, że środowisko zaawansowanej analizy powinno być kompletne i umożliwiać:

- dostęp do różnych źródeł danych wliczając w to tekst, logi, dane z sensorów, dane lokalizacyjne,
- przygotowanie, eksploracje i wizualizacje danych,
- rozwój i budowanie modeli analitycznych (na przykład modeli predykcyjnych),
- wdrażanie modeli i ich integracje z procesami i aplikacjami biznesowymi działającymi w firmie,
- wysoką wydajność i skalowalność zarówno dla rozwoju jak i wdrażania rozwiązań.

W czołówce rankingu znajduje się firma SAS, w [30] podkreślone zostaje, że posiada ona najszerszy zakres produktów zaawansowanej analizy. Ponad to, aplikacje takie jak Visual Analytics, Visual Statistics są ciągle udoskonalane, a obsługa jest coraz prostsza i bardziej intuicyjna. Klienci cenią w rozwiązaniach SAS jakość i elastyczność, co przekłada się na duże grono lojalnych klientów. Jedną z wad platformy SAS jest pokrywanie się funkcjonalności w produktach, przez co potencjalny klient może mieć problem w wyborze najlepszej opcji. Co więcej, dosyć duży odsetek użytkowników zgłaszał problemy z instalacją i aktualizacją oprogramowania. Gartner zwraca także uwagę, na drogi i nietransparentny system licencjonowania, co może być przyczyną wybierania przez organizacje innych rozwiązań.

Liderem w zakresie zaawansowanej analizy jest także firma IBM. Innowacyjna platforma *Watson*, która łączy przetwarzanie języka naturalnego i uczenie maszynowe aby przetwarzać ogromne zbiory nieustrukturyzowanych danych na wiedzę miała wpływ na to, że o firmie zrobiło się głośniej na rynku zaawansowanej analizy. Ponad to, oprogramowanie do analizy predykcyjnej SPSS jest dobrym produktem, który zapewnia duże możliwości, bez konieczności pisania kodu. Na pozycję firmy wpływa także jakość produktów i doświadczenie w różnych branżach. W raporcie zostaje zwrócona uwaga, na kiepską integracje produktów firmy, a także na to, iż konsultanci nie potrafią często dobrać odpowiedniego produktu dla klienta. Dodatkowo, słabe oceny wśród użytkowników zbierała także dokumentacja i mała ilość dostępnych szkoleń.

Rys. 3. Magic Quadrant dla platform zaawansowanej analizy [21]

Firma Microsoft i jej platforma zaawansowanej analizy osiągnęła najwyższy wskaźnik *Completness of Vision,* wpływ na to miało rozwiązanie *Cortana Analysis Gallery (CAG,* obecnie Cortana Intelligence Gallery *)*, które jest sklepem z rozwiązaniami analitycznymi dla platformy Azure. Istotne znaczenie ma też fakt, iż największy odsetek ankietowanych wybrał firmę Microsoft do własnych rozwiązań w przyszłości na podstawie harmonogramu(ang. *roadmap*). Klienci zwracają także uwagę na dobrą integracje z oprogramowaniem *open-source* takim jak R(język i platforma do obliczeń statystycznych) czy Python(język skryptowy). Dla niektórych respondentów, wadą rozwiązania typu CAG jest konieczność przechowywania danych w chmurze.

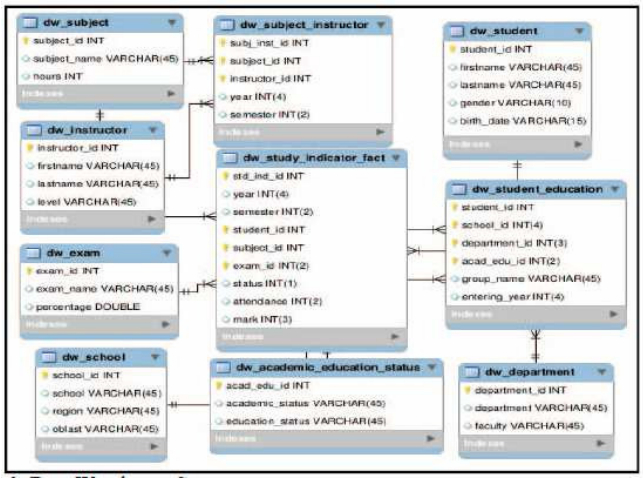
Na podstawie analizowanego raportu widoczny jest ciągły rozwój w zakresie zaawansowanej analizy. Istotny jest fakt, że organizacje wytwarzające oprogramowanie do tego celu mają różne wizje dotyczące przyszłości tego zagadnienia(Microsoft Cortana Intelligence Gallery, IBM WATSON). Widoczny jest także trend ograniczania konieczności pisania kodu do minimum i tworzenia intuicyjnych interfejsów użytkownika.

4.3. Możliwości integracji Big Data i rozwiązań Business Intelligence

Organizacje, posiadające dane typu Big Data i typu Business Intelligence mogą czerpać potencjalne korzyści z ich integracji. Jak zostało wspomniane w rozdziale 4.1, oba te rodzaje różnią się i mogą służyć do nieco innych celów. Dane typu Business Intelligence są ustrukturyzowane, oczyszczone i składowane w hurtowni danych, która jest zorientowana tematycznie. Dodatkowo, dane te pochodzą z pewnych źródeł i gromadzone są bardzo często latami, co za tym idzie, użytkownik, po wykonaniu zapytania do hurtowni danych otrzymuje odpowiedź, na podstawie wysokiej jakości danych [12]. Ze względu na charakterystykę Big Data, dane te nie zawsze muszą być wysokiej jakości, a także mogą pochodzić z różnych źródeł i nie być odpowiednio oczyszczone, co może przyczynić się do większego prawdopodobieństwa wyciągnięcia nieprawdziwych wniosków. Jednym z rozwiązań tego problemu jest utworzenie hurtowni danych, która będzie przechowywać zarówno dane typu BD jaki i BI. Takie podejście zaprezentowano w poprzednim rozdziale na rysunku 3.4. Qin H., Zhao Y., Qian Z., w artykule *On the Research of Data Warehouse in Big Data* [15] wymieniają potencjalne wyzwania, przed jakimi stoją projektanci takiej analitycznej bazy danych, są nimi:

- architektura danych i ich przetwarzanie,
- struktura bazy danych i infrastruktura systemu.

W publikacji *Data Warehouse on Hadoop Platform for Decision Support Systems in Education* [2] został przestawiony przykład projektu i implementacji hurtowni danych z wykorzystaniem technologii platformy Hive. Umożliwia ona budowę hurtowni danych w środowisku Hadoop. Hive oferuje struktury znane z relacyjnych baz danych takie jak tabele, kolumny i partycje wspiera także indeksowanie i metody optymalizacyjne, z uwagi na HDFS (rozproszony system plików platformy Hadoop), zapytania wykonywane są w sposób równoległy [2]. Projekt przedstawiony w artukule *Data Warehouse on Hadoop Platform for Decision Support Systems in Education* [2] miał na celu przeniesienie danych z bazy relacyjnej opartej o MySql do hurtowni danych zaimplementowanej z wykorzystaniem Hive. Dane pochodziły z uniwesytetu International Ataturk Alatoo University i dotyczyły edukacji na tejże uczelni. Do załadowania danych skorzystano z narzędzia Apache Sqoop, które zgodnie z dokumentem *Apache Scoop* [14], zastało zaprojektowane do transferu dużych ilości danych między bazami relacyjnymi, a platformą Hadoop. Schemat hurtowni danych wykorzystanej do projektu przedstawiony został na rysunku 3.



Rys. 4. Schemat hurtowni danych opartej na platformie Hive [2]

Autorzy projektu twierdzą, że spełnił on postawione na początku wymagania. Możliwa jest:

- wizualizacja danych,

- składowanie danych różnych typów w jednym miejscu,
- wykorzystanie danych do podejmowania decyzji.

Innym, mniej skomplikowanym podejściem jest wykorzystanie narzędzi Big Data Analytics i utworzenie różnych źródeł danych, analiza tych danych i przedstawienie ich na wspólnych kokpitach menedżerskich. To rozwiązanie można zastosować w przypadku posiadaniu hurtowni danych i baz NoSQL, gdy konieczne jest zestawienie takich danych na wykresach. Należy jednak pamiętać, że w tym wypadku znacznie ograniczane są możliwości analizy w odniesieniu do integracji we wspólnej hurtowni danych.

Potencjalne korzyści jakie niesie za sobą integracja danych typu BI i BD to przede wszystkim lepsze wspieranie podejmowania decyzji, gdyż opiera się ono nie tylko na niezweryfikowanych danych nieustrukturyzowanych, ale także na tych ustrukturyzowanych, gromadzonych latami z systemów ERP i CRM. Możliwe jest wykorzystanie prostych w obsłudze narzędzi Self Servive BI do tworzenia raportów i kokpitów menedżerskich przez kadrę zarządzającą, a także skorzystanie z platform zaawansowanej analizy obsługiwanych przez z analityków danych. Takie rozwiązanie, to niemalże wszystkie zalety systemów BI i BD dla organizacji takie jak: redukcja kosztów, redukcja czasu, poznanie klienta, optymalizacja produktu.

Propozycja integracji danych typu Business Intelligence i Big data jest rozwiązaniem innowacyjnym, gdyż na ryku nie ma gotowych platform, które oferują takie możliwości. Co za tym idzie, konieczne jest indywidualne określenie potrzeb klienta i decyzja projektantów o połączeniu tych dwóch systemów, a ostatecznie implementacja. Istotny fakt odgrywają potencjalne problemy przy integracji takie jak dobór odpowiedniej Bazy danych, czy narzędzi do budowy procesu ETL.

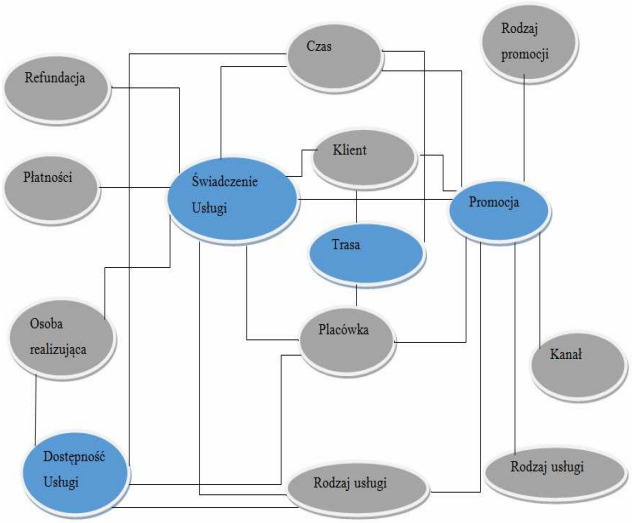
5. Studium przypadku - wdrożenie Big Data Analytics do organizacji z sektora medycznego

Autorski projekt systemu został wykonany dla firmy świadczącej usługi medyczne w zakresie badań laboratoryjnych, konsultacji lekarskich, oferującej także zabiegi operacyjne, wraz z możliwością pobytu w szpitalu pacjenta. Omawiana organizacja ma kilkanaście oddziałów w całym kraju i w różnych budynkach, niektóre z nich oferują wykorzystanie małych urządzeń działających w technologii Bluetooth w celu nawigacji wewnątrz budynku (Beacon) [15]. Takie rozwiązanie może posłużyć także do zbierania danych o drodze jaką przebył klient od wejścia, do punktu docelowego. Firma posiada własny system klasy CRM, który przechowuje dane na temat wizyt i pacjentów. Głównym celem wdrożenia jest zwiększenie rentowności poszczególnych usług medycznych o 20% w skali roku. Cele szczegółowe to:

- Zwiększenie rentowności funkcjonowania poszczególnych placówek medycznych o 20% w skali roku.
- Opracowanie strategii marketingowej promującej usługi medyczne.
- Wsparcie budowy regionalnych sieci przychodni.
- Wspieranie budowy strategii zwiększania dostępności usług medycznych.
- Opracowanie odpowiedniego oznaczenia drogi do recepcji w wynajmowanych budynkach.

5.1. Projekt systemu

Autorskie rozwiązanie, przygotowane na potrzeby niniejszej pracy składać się będzie z hurtowni danych, której schemat konceptualny przedstawiony został na rysunku 4, a miary i hierarchie wymiarów w tabeli 1 i 2. Dane pochodzić będą z istniejącego już systemu CRM, a także z urządzeń Beacon. Wizualizacja danych z hurtowni będzie odbywać się na raportach i kokpitach menedżerskich.



Rys. 5. Schemat konceptualny hurtowni danych -opracowanie własne

Tabela 1. Miary opisujące działalność organizacji - opracowanie własne

Świadczenie usługi medycznej	Cena jaką płaci klient
	Koszty jakie ponosi firma
	Zysk
	Jakość usługi - ocena klienta (0-5)
Promocja	Koszt promocji
	Liczba klientów nowych
	Liczba klientów stałych
	Wartość usług zrealizowanych w ramach promocji
	Zysk z promocji
Dostępność usługi	Liczba przewidywanych dni oczekiwania
	Liczba osób które zrezygnowały
Trasa	Czas przebycia trasy
	Liczba miniętych urządzeń Beacon

Oprócz klasycznej hurtowni danych, rozwiązanie przechowywać będzie dane z urządzeń Beacon w nierelacyjnej bazie *MongoDB*, która umożliwia przechowywanie dużych ilości nieustrukturyzowanych danych. Zasilanie hurtowni danych z systemu CRM i bazy MongoDB odbywać się będzie raz dziennie, w momencie najmniejszego obciążenia systemu. Za pomocą aplikacji typu Self Service BI możliwe jest utworzenie kokpitów menedżerskich które będą wizualizować odpowiednie dane i tym samym wspomagać podejmowanie decyzji w taki sposób, aby spełnione zostały określone na początku cele.

Tabela 2. Hierarchie wymiarów wykorzystane w projekcie - opracowanie własne

Klient	Województwo -> Miasto->Dzielnica ->Klient
Rodzaj usługi	Rodzaj usługi-> Podrodzaj usługi-> Usługa Specjalizacja usługi->Podspecjalizacja
Osoba Realizująca	Rodzaj wykonywanej pracy -> Typ personelu -> Nazwa stanowiska->Osoba Rodzaj Etatu -> Wymiar godzin
Płatności	Rodzaj - > Podrodzaj->Płatność
Placówka	Województwo -> Miasto->Dzielnica->Placówka
Czas	Rok->Kwartał->Miesiąc->Tydzień->Dzień->Godzina
Kanał Promocji	Rodzaj kanału-> Podrodzaj kanału -> Nazwa kanału
Rodzaje promocji	Rodzaj -> Podrodzaj->Promocja
Refundacja	Rodzaj->Podrodzaj->Refundacja
Lokalizacja	Województwo -> Miasto->Dzielnica

6. Wnioski

Głównym celem niniejszej pracy było określenie możliwości integracji danych typu Business Intelligence i Big Data. Cel ten został osiągnięty i na podstawie przeprowadzonych studiów literaturowych wyznaczono takie metody integracji jak:

- wspólna hurtownia danych,
- podłączenie narzędzia typu Big Data Analytics do źródeł danych BD i BI.

Na podstawie metod badawczych stwierdzono, że korzyścią z połączenia BD i BI będzie wiedza oparta na wysokiej jakości danych BIi wysokiej różnorodności BD. Zrealizowany zakres pracy wpłynie na kierunek badań w tym temacie, gdyż jak wynika z przeprowadzonego przeglądu literaturowego, nie jest on jeszcze popularny w środowisku naukowym. Co więcej, autorski projekt hurtowni danych integrującej w sobie dane Big Data i dane z systemu transakcyjnego może przynieść potencjalne korzyści w postaci zwiększenia rentowności usług medycznych w organizacji, która to rozwiązanie wdroży.

Istotnym wnioskiem jest także fakt, że producenci narzędzi do wizualizacji danych przykładają obecnie dużą uwagę do intuicyjności ich obsługi, przez co kadra zarządzająca potrzebuje mniejszego wsparcia ze strony działów IT. To przyczynia się do skrócenia czasu konstruowania kokpitów menedżerskich i raportów, a w konsekwencji szybszego i trafniejszego podejmowania decyzji. Ponad to, z przeanalizowanych raportów wynika, iż firmy użytkujące systemy klasy BI są zadowolone z korzyści jakie przyniósł firmie.

Literatura

[1] Arslan A. K., Colak C., Sarihan M. E., *Different medical data mining approaches based prediction of ischemic stroke,* computer methods and programs in biomedicine 130, 2016

[2] Bondarev A., Zakirov D., *Data Warehouse on Hadoop Platform for Decision Support Systems in Education*, 2015 Twelve International Conference on Electronics Computer and Computation (ICECCO)

[3] Devenport T.H., Dyche J., *Big data in big companies*

[4] Hannula M., Pirttimaki V., *Business intelligence empirical study on the top 50 Finnish companies*, Journal of American Academy of Business, Marzec 2003

[5] Kimball R., Ross M. *The Data Warehouse Toolkit Third Edition,*Wiley, 2013

[6] Mach-Król M., *ANALIZA I STRATEGIA BIG DATA W ORGANIZACJACH*, Studies & Proceedings of Polish Association for Knowledge Management 74, 2015

[7] Praca zbiorowa, *A review on application of data mining techniques to combat natural disasters*, Ain Shams Engineering Journal, 2016

[8] Praca zbiorowa, *Big Data for Dummies,* John Wiley & Sons, Inc. ,2013

[9] Praca zbiorowa, *Scalable real-time OLAP on cloud architectures,* J. Parallel Distrib. Comput. 79–80, 2015

[10] Romero O, Herrero V.,Abelló A.,Ferrarons J., *Turning Small Analytics on Big Data: Data Partitioning and secondary indexes in Hadoop ecosystem*, Information Systems 54, 2015

[11] Sarka D., Lah M., Jerkic G., *Exam 70-463:Implementing a Data Warehouse with Microsoft® SQL Server® 2012*, O'Reilly Media, Inc., 2012

[12] Stodder D., TDWI Best Practises Report - Business Driven Business Intelligence and Analytics. Achieving Value through Collaborative Business/IT Leadership, tdwi.ord, 3. kwartał 2014

[13] *Apache Scoop*, http://sqoop.apache.org/ [01.09.2016]

[14] *Beacony i mikrolokalizacja,* http://www.computerworld.pl/news/403124/Beacony.i.mikrolokalizacja.html [02.08.2016]

[15] *Big Data Analytics - What it is and why it matters*, http://www.sas.com/en_us/insights/analytics/big-data-analytics.html [01.07.2016]

[16] *Big Data - What it is and why it matters*, http://www.sas.com/en_th/insights/big-data/what-is-big-data.html [22.07.2016]

[17] *Connecting to Hadoop Hive,* http://kb.tableau.com/articles/knowledgebase/hadoop-hive-connection [16.08.2016]

[18] Data sources in Power BI Desktop, https://powerbi.microsoft.com/en-us/documentation/powerbi-desktop-data-sources/ [26.08.2016]

[19] *It Glossary*, http://www.gartner.com/it-glossary/ [21.07.2016]

[20] Magic Quadrant for Advanced Analytics Platforms, https://www.gartner.com/doc/reprints?id=1-2YEIILW&ct=160210&st=sb [01.07.2016]

[21] *Studium przypadku z wdrożenia systemu do budżetowania InForum BI Studio w Santander Consumer Bank*, http://prnews.pl/analizy/studium-przypadku-z-wdrozenia-systemu-do-budzetowania-inforum-bi-studio-w-santander-consumer-bank-3118696 [03.07.2016]

[22] Tableau & MongoDB: Visual Analytics on JSON at the Speed of Thought,http://www.tableau.com/about/blog/2015/6/tableau-mongodb-visual-analytics-json-speed-thought-39557 [05.08.2016]

[23] *Welcame to Apache Hadoop*, http://hadoop.apache.org/ [01.09.2016]

[24] *What is Big Data Analytics*, https://www01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html [07.07.2016]

[25] *What Is Data Mining,* https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm [29.07.2016]