Tomasz Kozyra

Zjawisko Big Data jako nowe światło dla analityki biznesowej

Ekonomiczne Problemy Usług nr 106, 199-213

2013

ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 781 EKONOMICZNE PROBLEMY USŁUG NR 106 2013

TOMASZ KOZYRA
Uniwersytet Warszawski

ZJAWISKO BIG DATA JAKO NOWE ŚWIATŁO DLA ANALITYKI BIZNESOWEJ

Wprowadzenie

Fenomenem o niewątpliwych cechach rozproszenia jest zjawisko noszące miano Big Data. Mówimy tu raczej o zjawisku, a nie o technologii, ponieważ wokół zjawiska tego koncentrują się nie tylko konkretne technologie, ale także kwestie zróżnicowania pochodzenia i typów danych, problemy zarządzania nimi, strategie organizacyjne oraz interesujące nas tu najbardziej – konsekwencje dla analityki biznesowej.

Era komunikacji elektronicznej przyczyniła się do lawinowego przyrostu danych. Często przywołuje się ogromny przyrost ilości blogów, tweetów czy e-maili, wykazując przy tym tempo wzrostu wolumenów danych w sieci globalnej. Anderson pokazuje jednak, że te media i kanały informacyjne nie stanowią o ogromie danych w Internecie. Według jego głośno wyrażanych opinii, sieć (WWW) należy uznać za skończoną, a przyszłość należy do Internetu[1]. Istotnie, Internet zazwyczaj kojarzony jest siecią WWW, jednak w jego obrębie duża część ruchu przypada na transmisje wideo, VoIP oraz P2P. Naturalnie pojawia się pytanie, na ile ta przeważająca część globalnego wolumenu danych może zostać wykorzystana przez organizacje, zwłaszcza

[1] Por. C. Anderson, *Who's to Blame*, „Wired Magazine" 2010, No. 18 (9), s. 36–37.

że dwa ostatnie z wymienionych kanałów mają charakter na poły prywatny lub ściśle prywatny.

W obecnej sytuacji świadomość ogromu danych jest coraz powszechniejsza. Inną jednak sprawą jest fakt generowania i gromadzenia danych, a inną możliwości ich interpretowania i analizowania. Od czasów powstania ery cyfrowej ilość danych zawsze wykraczała poza możliwości ich analizowania. Świadczyć o tym może choćby fakt, że przyrost danych przekracza równoległy przyrost mocy obliczeniowej procesorów opisywany przez prawo Moore'a[2]. O ile jednak prawo Moore'a wiąże się z ograniczeniami natury fizycznej (już na obecnym etapie pojawiają się głosy o załamaniu się tego prawa[3]), o tyle niepisane prawo przyrostu informacji ograniczone będzie głównie przez ograniczoność zasobów składowania, tj. przez przestrzeń konieczną do ich składowania, a nie ograniczenia wynikające z praw fizyki odniesionych do obecnie używanych materiałów stosowanych w jednostkach obliczeniowych. Powstaje zatem pytanie, jakie strategie informacyjne i organizacyjne należy przyjąć, aby dostępne dane w ich różnorodności i potężnej skali mogły służyć podejmowaniu decyzji?

Można przypuszczać, że wartość informacji maleje wraz z jej powszechnością. Stoi to w opozycji do jednej z tez Clevelanda, zgodnie z którą wartość informacji zależy od jej rozszerzania, czyli dzięki jej powszechności[4]. Cleveland poszukuje wartości, jaka płynie dla społeczeństwa, a jego rozważania nie koncentrują się w tym przypadku na wartości informacji, dzięki której można osiągnąć przewagę konkurencyjną. Można jego tezę odnieść do wnętrza organizacji, ponieważ wątpliwe wydaje się, aby radykalne ograniczanie dostępu do gromadzonych informacji wewnątrz organizacji miało pozytywny wydźwięk ekonomiczny. Na gruncie rynkowym przyswojenie przez organizację unikatowej i jej tylko dostępnej informacji może przyczynić się do wzrostu przewagi konkurencyjnej. Jednak w przestrzeni publicznej większość informacji jest

[2] Fayyad i Uthurusamy określają to prawo mianem prawa składowania (*Storage Law*), choć nie określają ścisłych wytycznych do jego weryfikacji, co ma miejsce w przypadku prawa Moore'a. Wskazują jednak na istotny trend, który potwierdza się w obecnym okresie. Por. U. Fayyad, R. Uthurusamy, *Evolving data into mining solutions for insights*, „Communications of the ACM" 2002, No. 45 (8), s. 28–31.

[3] S. Gaudin, *Physicist says Moore's Law is 'collapsing'*, „Computerworld" 2012, May 2.

[4] Por. H. Cleveland, *Information as a resource*, „Futurist" 1982, Dec., s. 36–37. W świetle współczesnych tendencji wiążących się z otwartością (np. modele biznesowe uwzględniające oferowanie oprogramowania *open source*) kwestia powszechności vs. prywatności informacji zyskuje nowe światło i nie jest tak spolaryzowana, jak to ukazano.

powszechnie dostępna, co oznacza, że potencjalnie każda organizacja może ją wykorzystać w celu osiągania przewagi konkurencyjnej. Zatem w chwili gdy dwie organizacje będą w posiadaniu tej samej, ogólnodostępnej informacji, osiągnięcie przewagi może być wątpliwe. Powstaje zatem pytanie, czy możliwe jest przekroczenie bariery powszechności zmniejszającej wartość informacji i wykorzystanie powszechności w celu zdobycia przewagi na rynku?

Informację tworzymy dzięki zinterpretowaniu danych. Sama interpretacja zależy od kontekstu, w którym lokujemy wykorzystywane dane. Otóż to właśnie, dzięki kontekstowi powszechnie zaistniałe informacje możemy ponownie przetworzyć i potraktować jako dane, które w kolejnym etapie podlegają nowym procesom interpretacji. Dzięki przetworzeniu ogólnodostępnej informacji w dane oraz ich nowej interpretacji powszechność informacji nie stanowi ograniczenia dla niesienia przez nią wartości oraz osiągania przewagi konkurencyjnej.

1. Charakterystyka Big Data

W świetle badań przeprowadzonych przez MIT Center for Digital Business organizacje, które określają siebie jako ukierunkowane przez dane, osiągają lepsze wyniki finansowe i operacyjne[5]. Organizacje nakierowane na podejmowanie decyzji popartych danymi są potencjalnie lepiej przygotowane do trudnego zadania tworzenia informacji z ogólnodostępnych rozproszonych danych. Mają zatem szanse na szybsze i efektywniejsze wykorzystanie potencjału zjawiska Big Data. Jak zatem powinniśmy ujmować ten fenomen? Spróbujmy przywołać niektóre definicje. Przez termin Big Data rozumie się:

- „techniki i technologie, które czynią operacje na danych o ogromnej skali dostępnymi ekonomicznie”[6];
- „nową generację technologii i architektur zaprojektowanych do uzyskania ekonomicznej korzyści z dużych wolumenów zróżnicowanych danych dzięki szybkiemu rejestrowaniu, odkrywaniu i/lub analizie”[7];

[5] E. Brynjolfsson, L. Hitt, H.H. Kim, *Strength in Numbers: How does data-driven decision-making affect firm performance?*, w: *ICIS 2011 Proceedings* 2011, Dec 6, Paper 13.

[6] B. Hopkins, B. Evelson, *Expand Your Digital Horizon with Big Data*, Forrester Research, Inc., 2011, Sept 30, s. 4.

[7] C. Olofson, D. Vesset, *Big Data: Trends, Strategies, and SAP Technology*, „Technical Report” 2012, No. 236135, IDC, s. 4.

– „dane, których rozmiar zmusza nas do spojrzenia poza wypróbowane metody [operowania danymi – T.K.], jakie dostępne są w teraźniejszym czasie”[8].

Ostatnia z definicji ukazuje historyczny rys powiązany ze zjawiskiem Big Data. Już w latach 80. ubiegłego wieku zbiory danych były na tyle duże, że wymagały zmiany tysięcy taśm w zrobotyzowany sposób. W latach 90. oznaczało to wykorzystanie raczej aplikacji Unixowych aniżeli aplikacji PC. W końcu współcześnie oznacza to być może porzucenie relacyjnego modelu baz danych oraz wykorzystanie systemów rozproszonych i przetwarzania równoległego. Zauważmy, że są to kolejne przykłady na ograniczenia związane z operowaniem danymi oraz ewentualnym wykorzystaniem ich na potrzeby przeprowadzanych analiz.

Z przywołanych definicji powinniśmy wyróżnić następujące elementy: ekonomię, dane oraz technologie (składowania, analiz itp.) z nimi powiązane.

Ekonomicznie uwarunkowana dostępność. Zarówno wdrożenie repozytoriów Big Data, jak i przenoszenie do nich danych można uznać za atrakcyjne z ekonomicznego punktu widzenia. Analizy dokonywane przez ekspertów od danych nie muszą opierać się na czasochłonnym procesie integrowania, oczyszczania i agregowania danych. Co więcej, dla specjalistów tych dostępne są środowiska przetwarzania równoległego, w których mogą wykorzystywać na przykład techniki drążenia danych na znacznych wolumenach danych przy użyciu transformacji wymagających niewielkich skryptów[9]. Duże znaczenie dla kontekstu ekonomicznego ma także możliwość implementacji środowisk analitycznych Big Data w środowiskach chmur obliczeniowych.

W odróżnieniu od tradycyjnych środowisk analitycznych bazujących na systemach klasy BI analityka Big Data nie musi podlegać procesom integracji, a jedynie pożądanej dla analizy transformacji danych. Ten aspekt ma kolosalne znaczenie z punktu widzenia czasu koniecznego do przeprowadzenia unikatowych analiz, gdyż etap integracji zostaje pominięty. Pominięcie etapu integracji bez wątpienia może przyczynić się do ograniczenia kosztów związanych z analityką. Taka strategia wiąże się z możliwością braku absolutnej precyzji, która jednak dla wielu problemów decyzyjnych nie ma krytycznego znaczenia.

[8] A. Jacobs, *The pathologies of big data*, „Communications of the ACM” 2009, No. 52 (8), s. 44.

[9] B. Hopkins, B. Evelson, *op. cit.*, s. 6.

Dane. Charakterystyka fenomenu Big Data bez wątpienia powinna być uzależniona od właściwego spojrzenia na specyfikę samych danych. Typowo przypisuje się im trzy kluczowe cechy (3V): skala (*volume*), szybkość (*velocity*) i różnorodność (*variety*):

Skala. Przywołany lawinowy wzrost danych jest głównym czynnikiem charakteru zjawiska Big Data. Skala jednak Big Data nie zależy jedynie od danych globalnej sieci. Coraz częściej to same organizacje generują dane, których wolumeny wykraczają poza możliwości składowania i analizy w tradycyjnych architekturach bazodanowych i aplikacjach analitycznych.

Szybkość. Dane przemieszczają się z coraz większą prędkością. Ten fakt ma ogromne znaczenie choćby w przypadku usług i rynków finansowych. Szybkość przemieszczania się danych oraz szybkość i częstotliwość rejestrowania zdarzeń (np. o charakterze lokalizacyjnym) mogą być skutecznie wykorzystywane nie tylko przez brokerów finansowych, ale także przez sieci handlowe, firmy telekomunikacyjne, gdzie właściwa i szybka reakcja na zmiany ma coraz większe znaczenie.

Różnorodność. Różnorodność wskazuje tu nie tylko na zmienność pochodzenia danych, ale także na ich znaczenie. Oznacza także różnorodność ich formatów: od plików tekstowych po dane generowane przez sensory.

Technologie. Analityka biznesowa zyskuje nowe światło, które jeszcze dekadę temu nie było w zasięgu rozwiązań systemów BI. Można wysnuć przypuszczenie, że wpłynęła na to nie tylko sama ilość i różnorodność danych, ale także nowe możliwości ich składowania i przetwarzania.

Na horyzoncie oferowanych rozwiązań składowania danych oprócz produktów opartych na klasycznych relacyjnych bazach danych pojawiło się wiele rozwiązań określanych mianem „NoSQL". Rozwiązania NoSQL cechują się między innymi możliwością obsługi dużych wolumenów danych i szybkością rejestrowania transakcji, rozproszoną architekturą oraz obsługą nieustrukturyzowanych danych. Do modeli NoSQL można zaliczyć bazy klucz–wartość, bazy dokumentowe, grafowe oraz kolumnowe.

Nie każdy rodzaj danych lub ich ilość czy zróżnicowanie pasują do jednego modelu infrastruktury bazodanowej. Na przykład dokumenty kodowane przy użyciu XML mogą być przechowywane w bazach XML, a relacje w sieciach społecznościowych z natury są grafami i bardziej pasują do bazy grafowej. Jeżeli organizacja będzie stosować wiele rożnych modeli składowania danych,

co skutkuje dodatkowym rozproszeniem systemów, to staje między innymi przed problemem integracji heterogenicznych modeli (np. problem dostępu do platform NoSQL przez standardowe aplikacje BI).

W chwili gdy konwencjonalne infrastruktury baz relacyjnych nie mogą obsłużyć znaczących wolumenów danych, wśród głównych możliwości technologicznych należy wymienić architekturę MPP (*massively parallel processing architecutre*) oraz rozwiązania oparte na platformie Hadoop. Decyzja o wyborze jednej z tych opcji zależy gównie od stopnia różnorodności danych oraz konieczności zmienności schematów danych: hurtownie danych bazujące na architekturze MPP oparte są na definiowanych schematach, podczas gdy Hadoop nie stawia ograniczeń co do struktury danych.

2. Analityka Big Data a tradycyjna analityka biznesowa

W świetle powyższej charakterystyki może powstać pytanie, czy analityka Big Data nie jest po prostu analityką, która może być przyrównana do tradycyjnych rozwiązań analityki biznesowej, takich jak *business intelligence*.

Bez wątpienia posiadanie dużych ilości danych, obecne już w początkach rozwoju sieci WWW czy elektronicznego handlu, musi podlegać właściwym metodom analizy powiązanym z terminem „analityka odkrywająca". Realizowana może być ona przy użyciu narzędzi bazujących na SQL, drążeniu danych, analizie statystycznej, wizualizacji danych, przetwarzaniu języka naturalnego oraz tekstu itp. Analitycy Big Data próbują odkrywać nowe, nieznane dotychczas fakty, zaś typowe rozwiązania BI związane są w większości przypadków z raportowaniem tego, co wiemy o niewiadomych:

„Hurtownie danych i narzędzia BI świetnie odpowiadają na powtarzane w koło pytania, takie jak: »jaka była sprzedaż Marii w tym kwartale?«. Jednak gorzej sprawdzają się w badawczych, nieprzewidywalnych pytaniach »co – jeżeli«, które mają znaczenie dla planowania i podejmowania decyzji, ponieważ szybka eksploracja niestrukturyzowanych danych jest typowo trudna do przeprowadzenia, a tym samym kosztowna"[10].

[10] A. Croll, *Three Kinds of Big Data* w *Big Data Now*, 2012 Edition, O'Reilly Media, Sebastopol 2012, s. 60–61.

Funkcjonowanie systemów przetwarzających dane źródłowe opiera się na swoistej zasadzie ograniczania wolumenu danych, ponieważ część danych w procesie oczyszczania jest tracona. Dane z obszaru Big Data powinny podlegać zasadzie opierającej na składowaniu danych, a ściślej dyktacie utrzymywania wszelkich danych, o ile to możliwe. Przeciwdziałanie utracie danych może przyczynić się do trafniejszych decyzji. Z punktu widzenia analityki skala Big Data zapewnia gigantyczne próbki statystyczne, co niewątpliwie usprawnia takie metody analiz, jak drążenie danych czy analizy statystyczne. Ponadto techniki zaawansowanej analizy Big Data są relatywnie sprawne w uzyskiwaniu pożądanych rezultatów z nieprzetworzonych danych źródłowych, z danych o niskiej jakości, danych niestandardowych. W tym kontekście dobrym przykładem jest wykrywanie oszustw przy użyciu analityki Big Data. Mają tu bowiem znaczenie wszelkie dane będące sygnałem odstępstwa, które może być przeoczone w chwili stosowania tradycyjnych metod oczyszczania danych i procesów ETL, jeśli wykorzystywane są w analityce Big Data.

Powszechność informacji oraz zaawansowane możliwości analizy i interpretacji wymuszają podejmowanie określonych działań, mających na celu efektywne wykorzystanie zjawiska Big Data przez organizacje. Barton i Court wymieniają trzy główne typy strategii, które firmy powinny opanować, aby wykorzystać potencjał Big Data[11]:

1. **Wybór właściwych danych.** Różnorodność danych jako cecha Big Data wymusza sprawny dobór źródeł danych. Właściwe zadanie decydentów i analityków odpowiedzialnych za formowanie infrastruktury i jej wykorzystanie polega na dopasowaniu źródeł danych do konkretnych problemów biznesowych lub do pojawiających się szans rozwoju przedsięwzięć. Koniecznym warunkiem osiągania wartości z określonych typów danych jest uzyskanie wsparcia działów IT, gdyż infrastruktura IT często nie jest przystosowana do integracji wielu typów danych, a zwłaszcza danych niestrukturyzowanych. Zainteresowane strony mogą wykorzystać taktykę krótkoterminową, dzięki której identyfikowane i łączone są najważniejsze dane.

2. **Tworzenie modeli przewidujących i optymalizujących wyniki.** Najprawdopodobniej najefektywniejsza strategia dotycząca Big Data nie zaczyna się od danych, ale raczej od określenia, jak model będzie wpływać

[11] D. Barton, D. Court, *Making Advanced Analytics Work for You*, „Harvard Business Review" 2012, No. 90 (10), s. 80–83.

na poprawienie wyników. Konieczne jest przy tym zachowanie względnej prostoty modelu, gdyż jego złożoność wpływa na wydajność systemów.

3. **Przemiany organizacyjne.** Po pierwsze, analityka powinna być zorientowana na codzienne działania i procesy do nich dopasowane. Oznaczać to może choćby ustanowienie zespołu zadaniowego do spraw analityki, który na przykład podczas spotkań z menedżerami odpowiedzialnymi za politykę cenową i promocję może precyzyjnie określać typy podejmowanych decyzji koniecznych do ustalania konkretnych polityk. Pozwolić to może na skuteczny dobór narzędzi, które wspierać mogłyby procesy decyzyjne. Po drugie, zaawansowane modele powinny być implementowane na potrzeby pracowników pierwszej linii. Po trzecie, organizacja powinna wzbogacać możliwości wykorzystywania modeli przez decydentów, co może być osiągane dzięki zapewnianiu szkoleń lub mierzeniu wpływu i wykorzystywania modeli wraz z ewentualnym promowaniem i wynagradzaniem praktyk mających na celu stosowanie wdrażanych praktyk analityki Big Data.

Zjawisko Big Data jest obecnie przedmiotem szerokiego zainteresowania. W celu wdrożenia jakiejkolwiek strategii powiązanej z tym zjawiskiem organizacje muszą odpowiedzieć na pytanie, czy potencjał analityczny tego zjawiska przyniesie pożądane korzyści. Zasadniczo można je osiągnąć w trzech głównych obszarach.

4. **Analityka klienta.** Era Internetu zmieniła sposób rozumienia zachowań klientów i właśnie dlatego duże sieci internetowe wypchnęły w wielu przypadkach firmy tradycyjne. Pośrednim tego przejawem jest przemiana tradycyjnego modelu marketingowego (4P) w modele uwzględniające dodatkowy komponent, a mianowicie ludzi, których działania i decyzje w erze cyfrowej mogą być w precyzyjny sposób analizowane:

„W chwili gdy klienci zaczęli dokonywać zakupów w sieci, zrozumienie ich zachowań znacząco wzrosło. Detaliści sieciowi mogą śledzić nie tylko to, co kupili klienci, ale także to, co obserwowali, jak poruszali się po sklepie, jak bardzo podatni byli na promocje, recenzje i układ stron, jakie wykazywali podobieństwa do innych kupujących lub ich grup. Natychmiast zaczęli tworzyć algorytmy przewidujące, jakie książki klienci chcieliby przeczytać – algorytmy, które poprawiają swe działanie za każdym razem, gdy klient odpowiada na rekomendację lub ją ignoruje. Tradycyjni detaliści po prostu nie mają dostępu do tego typu informacji, nie mówiąc o dostępie do nich we

właściwym czasie. Nie dziwi zatem, że Amazon zmusił wiele tradycyjnych firm (*brick-and-mortar*) do wycofania się z rynku"[12].

Z punktu widzenia danych transakcyjnych pojedynczy wpis w serwisie społecznościowym nie niesie tyle informacji co konkretny zapis transakcji. Jednak wartość informacyjna tych dwóch różnych źródeł (danych transakcyjnych i nietransakcyjnych) wzrasta po ich powiązaniu. Na przykład specjalnie filtrowane ogromne ilości komentarzy mogą być powiązane z historią zakupów danego produktu lub kampanii sprzedażowej.

Do specyficznych zastosowań analityki Big Data w obszarze powiązanym z działaniami i zachowaniami klientów należy zaliczyć: wpływ zachowań w społecznościach na działania marketingowe, segmentację baz klientów, rozpoznawanie cech sprzedaży oraz szans rynkowych. Należy tu podkreślić, że analityka Big Data odniesiona do klientów wnosi istotną wiedzę, gdyż zarejestrowane fakty dotyczące transakcji lub preferencji nie dają odpowiedzi między innymi na pytanie „dlaczego?". W ogólnym ujęciu informacje uzyskiwane dzięki Big Data wykraczają „poznawczo" poza informacje o faktach – dostarczają wiedzy o tym, „co mogło się wydarzyć, co powinno się wydarzyć lub co się wydarzy"[13].

5. **Rozszerzenie potencjału BI**. Dzięki analizom predykcyjnym, drążeniu danych, analizie wieloczynnikowej czy wizualizacji danych analityka Big Data znacznie poszerza możliwości tradycyjnych środowisk BI. Analityka Big Data wydaje się naturalnym rozwinięciem tradycyjnych systemów BI i należy ją rozumieć jako dziedzinę szeroko rozumianej analityki biznesowej. Nie oznacza to jednak, że należy ją traktować jako niezależną materię analityczną. Przeciwnie, w wielu wypadkach konieczne jest całościowe spojrzenie na analizowane problemy, które uwzględnia nie tylko optykę Big Data, ale także zastane platformy analityczne. Różnorodność danych sprawia, że niektórym przypisuje się odrębne znaczenie. Tymczasem stanowią one część „informacyjnego kontinuum". Właśnie dlatego należy się spodziewać, że informacje pochodzące ze źródeł Big Data będą zasilać tradycyjne raporty czy kokpity menedżerskie, a wiele organizacji już stosuje takie rozwiązania. Ponadto część

[12] A. McAfee, E. Brynjolfsson, *Big Data: The Management Revolution*, „Harvard Business Review" 2012, No. 90 (10), s. 62.

[13] S. Swoyer, *Big Analytics: The Next Generation w Big Data Analytics*, TDWI E-book, TDWI Research Inc., September 2012, s. 9.

zastosowań analityki Big Data realizowana jest w przestrzeni korporacyjnych hurtowni danych (EDW), wpisując się ściśle w klasyczną architekturę systemów klasy BI.

6. **Zastosowania tematyczne**. Choć możliwość wykorzystania Big Data na polu analizy zachowań klientów jest podkreślana najczęściej, analityka Big Data to nie tylko analityka klienta. Istnieje szereg zastosowań, które nie są bezpośrednio powiązane z tym obszarem. Można do nich zaliczyć między innymi: optymalizację logistyki i operacji w łańcuch dostaw, kwantyfikację ryzyk oraz aplikacje mające na celu wykrywanie oszustw. Za istotną korzyść można uznać także możliwość automatyzacji procesów biznesowych w czasie rzeczywistym, na przykład podejmowanie decyzji o przyznaniu kredytu.

Pomimo znaczącego potencjału analityki Big Data wiąże się ona z istotnymi barierami, które można podzielić na organizacyjne i technologiczne. Już przy pierwszym spojrzeniu na zagadnienie Big Data może pojawić się wątpliwość, czy dana organizacja będzie w stanie podjąć się wyzwań powiązanych z Big Data. Jedną z najczęściej wymienianych barier jest bariera wiedzy oraz koniecznych zasobów ludzkich[14]. Istotnie, obszar analityki Big Data wymaga swoistych kompetencji, które znacząco wykraczają poza typową wiedzę przyswajaną przez analityków z obszaru *business intelligence*. Kompetencje te wiążą się choćby z umiejętnościami projektowania i tworzenia architektury Big Data oraz z zapewnieniem jej użyteczności dla użytkowników końcowych.

Rozwój zjawiska Big Data wpłynął na nowy obszar nazywany „nauką o danych" (*data science*). Obejmuje on swoim zakresem takie dziedziny, jak matematyka, programowanie, oraz wiąże się ze swoistym zmysłem naukowym. Dziedzina ta, która ma szczególne znaczenie dla analityki Big Data, doprowadziła do wyłonienia się nowych kompetencji osób odpowiedzialnych za przygotowywanie analiz i ich udostępnianie – ekspertów od danych (*data scientist*). Do cech charakteryzujących tę nową rolę w organizacjach należy zaliczyć: ekspercką wiedzę w konkretnej dziedzinie naukowej, zdolność do rozkładania problemów na zestaw możliwych do testowania hipotez, umiejętność efektywnego komunikowania rezultatów (skomplikowane analizy muszą zostać w odpowiedni sposób przyswojone przez osoby korzystające z rezultatów) oraz zdolność do spojrzenia na problemy z wielu kreatywnych stron.

[14] P. Russom, *Big Data Analytics*, TDWI Research Inc. 2011, s. 12.

Jedną z powszechnych barier dla analityki Big Data jest przywiązanie jej do zespołów odpowiedzialnych za tradycyjne środowisko hurtowni danych i *business intelligence*. Tymczasem, jak przekonuje Russom, większość zastosowań Big Data ma charakter ściśle powiązany z działaniem konkretnych departamentów. Pokrywa się to zresztą z wymienionymi obszarami zastosowań omawianego zjawiska. Ponadto wiele inicjatyw powiązanych z Big Data wychodzi od konkretnych działów, które w wielu przypadkach są w stanie zapewnić konieczne fundusze do ich realizacji. Właśnie dlatego szansą dla wielu projektów Big Data jest wsparcie ekspertów od danych, niekoniecznie związanych z korporacyjnym zespołem analitycznym.

Większość współczesnych środowisk analitycznych stanowi istotną barierę technologiczną dla realizacji inicjatyw powiązanych z Big Data. Jedną z głównych przeszkód natury technicznej jest skalowalność zastanych środowisk. Przy czym skalowalność oznacza tu niedopasowanie większości środowisk BI do skali wolumenów danych. Napotyka się także problemy szybkiego przetwarzania niejednokrotnie równoległych zapytań oraz ładowania danych. Ponadto wiele problemów może pojawić się, jeśli hurtownie danych modelowane są jedynie pod kątem raportów i OLAP. Te problemy mogą przyczynić się do decyzji o zmianie architektury platformy analitycznej. Według badań przeprowadzonych przez TDWI Research połowa firm nie planuje zmiany obecnych platform analitycznych[15]. Wiąże się to częstokroć z brakiem koniecznych funduszy, a także zaspokojeniem obecnych potrzeb w zakresie analityki biznesowej. Jedna trzecia badanych organizacji zamierza wymienić obecne platformy w ciągu trzech lat.

Całkowita wymiana platformy analitycznej nie zawsze jest możliwa czy pożądana. Jak widzieliśmy, naturalną drogą dla analityki Big Data są choćby inicjatywy mające na celu realizację środowisk wykorzystujących potencjał omawianego fenomenu na poziomie departamentów. W praktyce oznacza to stosowanie dwóch równoległych stylów analityki, co prowadzi do istotnych zależności między danymi z równoległych platform. Z jednej strony aplikacje Big Data mogą wykorzystywać na swoje potrzeby dodatkowe dane z hurtowni danych, z drugiej zaś w wielu przypadkach rezultaty analityki mogą zasilać hurtownie danych, a w konsekwencji platformę BI. Jeżeli zatem analityka Big Data ma stać się rozszerzeniem tradycyjnych platform BI, to technologie Big

[15] *Ibidem*, s. 20.

Data muszą być uzupełniane o narzędzia integracji danych łączące te dwa środowiska, o ile funkcjonować będą jako oddzielne platformy.

Decyzja o wdrożeniu odrębnych platform analitycznych (Big Data i BI) zależy między innymi od odpowiedzi na następujące pytanie: czy korporacyjna hurtownia danych będzie w stanie sprostać wymaganiom skali Big Data bez wpływu na jej dotychczasowe zadania, związane choćby z raportowaniem lub OLAP? Pytanie to ostatecznie sprowadza się do kwestii obsługi dostatecznej ilości zadań oraz związanego z nimi obciążenia. Czy będzie ona w stanie efektywnie przetwarzać dane o specyfice Big Data? Wiele implementacji korporacyjnych hurtowni danych spełnia takie oczekiwania, między innymi dzięki analityce *in-database*. Zdarza się jednak, że obciążenia związane z zarządzaniem Big Data i zaawansowaną analityką są na tyle znaczne, że analityka ta w obrębie korporacyjnej hurtowni danych jest nie do przyjęcia. Dodatkowym problemem natury technicznej jest stosowanie ETL w odniesieniu do Big Data (np. dane dzienników systemowych lub pochodzące z sensorów ładowane do baz analitycznych)[16]. Co więcej, ewolucja hurtowni danych nakreśliła ich typowy charakter:

„Istniejące korporacyjne hurtownie danych i bazy relacyjne celują w przetwa-rzaniu danych ustrukturyzowanych i mogą przetrzymywać ogromne zbiory danych, jednak za pewną cenę: wymóg ustrukturyzowania ogranicza możliwość przetwarzania pewnych typów danych i wprowadza inercję, która sprawia, że hurtownie danych są niedopasowane do sprawnej eksploracji heterogenicznych danych o dużej skali. Rozmiar wysiłku wkładanego w składowanie danych (w hurtowniach) sprawia, że wiele cennych źródeł danych nigdy nie jest analizowanych. Właśnie w tym aspekcie Hadoop może doskonale się sprawdzić”[17].

Platforma Hadoop jest obecnie jednym z najczęściej stosowanych rozwiązań zapewniających obsługę ogromnych wolumenów danych. Jej głównymi zaletami są skalowalność oraz możliwość przetwarzania równoległego. Te dwie zalety osiągane są głównie dzięki architekturze sprzętowej opartej na tanich, masowo tworzonych serwerach. Dzięki rozproszonemu systemowi plików (*Hadoop Distributed File System* – HDFS) Hadoop zapewnia obsługę danych zawartych w plikach, co zdecydowanie usprawnia proces ich przecho-

[16] A. Croll, *op. cit.*, s. 61.

[17] E. Dubmill, *What is Big Data?*, w: *Big Data Now*: 2012 *Edition*..., s. 10.

wywania. Przy zastosowaniu tradycyjnej bazy relacyjnej dane wymagałyby modelowania, integracji i ładowania. Dodatkowym komponentem platformy Hadoop jest framework MapReduce, który uzupełnia HDFS o możliwą analitykę. Hadoop na obecnym etapie wiąże się między innymi z takimi problemami, jak manualne kodowanie zadań MapReduce, wymagające niezbędnej wiedzy i umiejętności.

Właściwa realizacja platformy analitycznej Big Data zasadniczo może być realizowana na trzy sposoby: platforma software'owa, platforma sprzętowo-software'owa (np. dzięki rozwiązaniom data *warehouse appliance*) lub przy użyciu chmur obliczeniowych. Wybór jednej z tych możliwości podlega uwarunkowaniom ekonomicznym, regulacyjnym i technologicznym.

Wolumeny danych w analityce Big Data niejednokrotnie przekraczają możliwości ich przenoszenia (np. z systemów i baz zewnętrznych). Jest to typowy problem funkcjonowania danych w środowiskach rozproszonych, a zwłaszcza w środowiskach chmur obliczeniowych o charakterze hybrydowym. W chwili gdy wolumeny przenoszonych danych koniecznych do analizy przekroczą kryteria techniczne i ekonomiczne, organizacje mogą rozważyć przeniesienie aplikacji analitycznej do środowiska chmury obliczeniowej. Inna sytuacja dotyczy postulatu zerowego opóźnienia w przedsiębiorstwie i wiąże się także z transmisją danych. Tym razem jednak to nie wolumeny danych, ale opóźnienie w dostępie do nich ma znaczenie krytyczne. Dotyczy to głównie transakcyjnych systemów finansowych. W takich przypadkach realizacja aplikacji w środowisku bliskim źródła powstawania danych może mieć istotne uzasadnienie biznesowe (opóźnienie niektórych operacji o milisekundy może mieć negatywny wpływ na przewagę konkurencyjną).

Wspominaliśmy o ekonomicznych zaletach płynących z możliwości wdrożenia analityki Big Data w środowisku chmur obliczeniowych. Naturalnie dotyczy to skali inwestycji oraz ewentualnego jej przełożenia na korzystne wyniki. Nie wszystkie bowiem organizacje są w stanie ponieść znaczące nakłady konieczne do wdrożenia platformy analitycznej, np. przy użyciu takich rozwiązań, jak rozwiązania typu *software appliance*. Co więcej, niektóre projekty mogą wiązać się jedynie z tymczasową potrzebą.

Nawet jeśli platforma analityczna Big Data zdaje się być w zasięgu danej organizacji, z jej funkcjonowaniem wiążą się między innymi procesy wstępnego oczyszczania danych o silnie nieuporządkowanym charakterze. Zanim

zatem dojdzie do zastosowania konkretnych zbiorów danych i dokonania analiz, dane muszą przejść przez żmudny, wstępny proces przyswajania ich na potrzeby tych analiz. Szacuje się, że około 80% zadań związanych z danymi w rozwiązaniach Big Data polega na wstępnym oczyszczaniu lub jak ujmuje to Warden: „zamianie nieładu danych w coś użytecznego”[18]. W takich przypadkach organizacje mogą zrezygnować z tych procesów i korzystać z usług określanych mianem DaaS (*Data as a Service*). Usługi tego typu wprowadzają nowy sposób korzystania z potencjału zjawiska Big Data. Oferowane są na rynkach danych (*data marketplaces*), gdzie usługodawcy udostępniają zbiory danych, z których korzystać mogą organizacje. Naturalnie usługodawcy w modelu DaaS powinni zmierzać do udostępniania danych oczyszczonych i względnie przetworzonych, a jakość dostępnych wolumenów wpływać będzie na konkurencyjność w obrębie tych rynków.

Podsumowanie

Analityka Big Data stanowi nowy obszar analityki biznesowej i w skuteczny sposób może rozszerzyć potencjał platform *business intelligence*. Decyzja odnośnie do zastosowania analityki Big Data zależy od zdolności organizacji do realizacji platformy realizującej zadania tej analityki. W systemach rozproszonych bariery płynące z rozproszenia danych, ich heterogeniczności i szybkości przemieszczania się oraz przeszkody natury organizacyjnej mogą być łatwiej pokonywane i skutecznie wykorzystywane niemalże przez dowolną organizację, a ogrom Big Data nie musi już przerażać nawet najmniejszych podmiotów rynkowych.

Literatura

Anderson C., *Who's to Blame*, „Wired Magazine” 2012, No. 18 (9).

Brynjolfsson E., Hitt L., Kim H. H., *Strength in Numbers: How does data-driven decision-making affect firm performance?*, *ICIS 2011 Proceedings* 2011, Paper 13.

Barton D., Court D., *Making Advanced Analytics Work for You*, „Harvard Business Review” 2012, No. 90 (10).

[18] *Ibidem*, s. 9.

Cleveland H., *Information as a resource*, „Futurist", December 1982.
Croll A., *Three Kinds of Big Data* w *Big Data Now: 2012 Edition*, O'Reilly Media, 2012.
Dubmill E., *What is Big Data?* w *Big Data Now: 2012 Edition*, O'Reilly Media, 2012.
Fayyad U., Uthurusamy R., *Evolving data into mining solutions for insights*, „Communications of the ACM" 2002, No. 45 (8).
Gaudin S., *Physicist says Moore's Law is collapsing*, „Computerworld", 2 May 2012, http://www.computerworld.com/s/article/9226758/Physicist_says_-Moore_s_Law_is_collapsing_.
Hopkins B., Evelson B., *Expand Your Digital Horizon with Big Data*, Forrester Research, Inc., 30 September 2011.
Jacobs A., *The pathologies of big data*, „Communications of the ACM" 2009, No. 52(8).
Labrinidis A., Jagadish H.V., *Challenges and opportunities with big data*, „Proc. VLDB Endow" 2012, No. 5 (12).
McAfee A., Brynjolfsson E., *Big Data: The Management Revolution*, „Harvard Business Review" 2012, No. 90 (10).
Olofson C., Vesset D., *Big Data: Trends, Strategies, and SAP Technology*, IDC 2012, No. 236135.
Russom P., *Big Data Analytics*, TDWI Research Inc. 2012.
Swoyer S., *Big Analytics: The Next Generation w Big Data Analytics*, TDWI E-book, TDWI Research Inc., September 2012.

BIG DATA: A NEW PERSPECTIVE ON BUSINESS ANALYTICS

Summary

In contemporary IT systems rapid explosion of data is observed, which results in the emergence of new phenomenon called Big Data. In this article Big Data is evaluated across determinants such as economics and technology. The main differences between traditional analytical environments implemented in business intelligence platforms and Big Data analytics are discussed along with the aspects of their co-existence. The article also presents business benefits of Big Data analytics as well as challenges and opportunities of using this form of analytics in cloud computing.

Translated by Tomasz Kozyra