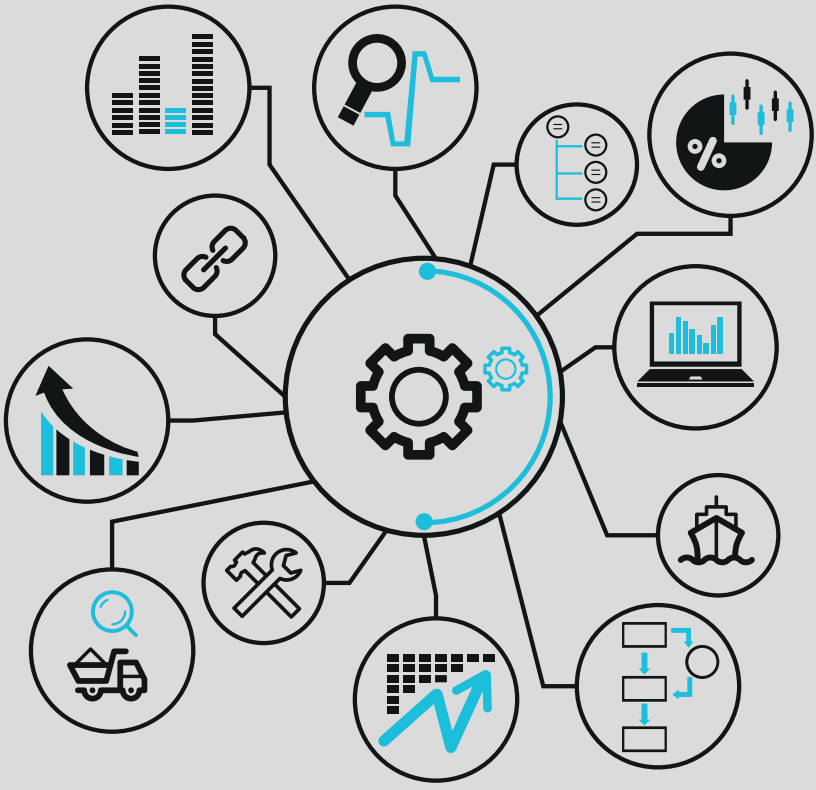
HYUNJOUNG LEE IL SOHN

Big Data w przemyśle

Jak wykorzystać analizę danych do optymalizacji kosztów procesów?



PWN

Dane oryginału:
Fundamentals of Big Data Network Analysis for Research and Industry

Z języka angielskiego tłumaczył: **WITKOM Witold Sikorski; Maciej Baranowski**

Projekt okładki i stron tytułowych **Piotr Fedorczyk**

Wydawca **Adam Filutowski**

Koordynator ds. redakcji **Renata Ziółkowska**

Redaktor **Małgorzata Dąbkowska-Kowalik**

Produkcja **Mariola Grzywacka**

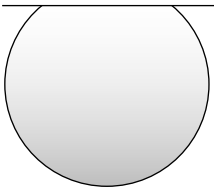
Skład i łamanie **Pracownia Obrazu – Anna Sandecka-Ląkocy**

ISBN: 978-83-01-18733-0

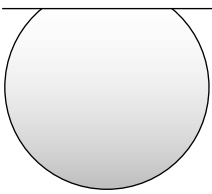
Wydanie I

Wydawnictwo Naukowe PWN SA
02-460 Warszawa, ul. Gottlieba Daimlera 2
tel. 22 69 54 321; faks 22 69 54 288
infolinia 801 33 33 88
e-mail: pwn@pwn.com.pl; reklama@pwn.pl
www.pwn.pl

Druk i oprawa: OSDW Azymut Sp. z o.o.



Spis treści



Wstęp do wydania polskiego

Rosnąca popularność metod analizy sieci społecznych i dostępność wielu programów, za pomocą których można z łatwością przygotować niezwykle atrakcyjne graficznie wizualizacje, mogą wywołać wrażenie, że interpretacja wielopoziomowych i zawiłych relacji między dowolnie wybranymi elementami sieci stanie się prostym w obsłudze, dostępnym każdemu narzędziem, dzięki któremu można szybko uzyskać rzetelne, miarodajne informacje. Nawet najdoskonalsze narzędzie nie jest jednak gwarancją uzyskania perfekcyjnych wyników i pełne jego wykorzystanie wymaga sporej wiedzy oraz umiejętności. Ponadto trzeba pamiętać, że skomplikowany, zaczerpnięty z wielu różnych dziedzin wiedzy (teoria grafów, socjologia, statystyka) aparat pojęciowy i ogromna swoboda w wyborze kryteriów badań otwierają także szerokie pole do różnego rodzaju nadużyć, manipulacji i uproszczeń. Z nieprzebranych zbiorów Big Data można przecież wyodrębnić arbitralnie określony podzbiór i stosując pasującą do tezy badania metodę, uzyskać czytelne, intuicyjnie – wydawałoby się – oczywiste rezultaty, a na ich podstawie podejmować ważne biznesowe decyzje (albo kogoś do nich przekonywać).

Autorzy tej książki podjęli się niezwykle ambitnego zadania: wyczerpującego, a jednocześnie maksymalnie zwięzłego przedstawienia twardych, naukowych podstaw SNA (Social Network Analysis – analiza sieci społecznych) na neutralnym, a zarazem świetnie sobie znanym terenie (przemysł stalowy), wykorzystując w tym celu powszechnie dostępne dane i przykłady z rzeczywiście wykonanych badań.

Ten biznesowy punkt widzenia uwypukla bezstronność metody, a oryginalne ujęcie problematyki, mimo że dotyczące konkretnej branży, będzie zapewne interesujące również dla polskiego odbiorcy.

Prezentując tak dużą dawkę teorii związanej z analizą sieci społecznych, wraz z praktycznymi przykładami, autorzy nie ustrzegli się jednak pomyłek i nieścisłości merytorycznych. Tłumacz, starając się zachować maksymalną wierność oryginałowi, a jednocześnie dbając o jak najlepszą jakość publikacji, sygnalizuje te wątpliwości za pomocą przypisów. W przypisach podano także informacje dotyczące aktualnych wersji opisywanych aplikacji.

1 Dlaczego Big Data?

Ilość danych jest ogromna i cały czas rośnie w szalonym tempie. Jednocześnie przybywa danych zbędnych, a wykonanie bardziej efektywnej, rzetelnej analizy wymaga ich przefiltrowania i usunięcia. Umiejętność wyodrębniania ze zbiorów danych prawidłowych i przydatnych informacji staje się czymś nieodzownym. Dzięki analizie Big Data przedsiębiorstwo zyskuje możliwość oddzielenia „ziarna od plew" i poszerzenia swojej początkowo dość wąskiej perspektywy. Istotą Big Data nie jest objętość (ilość) danych, szybkość ich przepływu ani różnorodność, lecz poszerzenie horyzontów myślowych oraz inne spojrzenie na dane. Chcesz zobaczyć cały las? To nie wychodź z niego, ale wespnij się na szczyt góry. Podobnie rzeczy mają się z Big Data. Szukasz istotnych informacji? Wzbij się niczym ptak w przestworza, a im wyżej się wniesiesz, tym szersze będzie twoje pole widzenia. Aby zobaczyć z zewnątrz to, czego się nie da uchwycić pozostając wewnątrz, potrzebny jest punkt widzenia obejmujący cały las. I tutaj właśnie wkracza Big Data.

1.1. Big Data

Zainteresowanie analizą Big Data jest coraz większe. Gartner, jedna z wiodących na świecie firm specjalizujących się w analizach rynku, wskazała Big Data jako jedną z dziesięciu strategicznych technologii [1] w latach 2012 i 2013; a w roku 2014 uznała Big Data i Actionable Analytics[1] za podstawowe technologie w strategii inteligentnego zarządzania [2]. Co więcej, w styczniu 2012 r. w Davos, na corocznym spotkaniu światowych liderów politycznych i ekonomicznych, którzy w ramach Światowego Forum Ekonomicznego [3] omawiają globalne problemy, wybrano Big Data jedną z dziesięciu

[1] Analiza ze wskazówkami do działania (przyp. tłum.).

kluczowych technologii dla przyszłego rozwoju. Ponieważ obecnie na pierwszym planie są próby wyjścia z kryzysu finansowego, jak również problemy związane ze zmianą klimatu, energią, ubóstwem i bezpieczeństwem, wskazanie Big Data zdaje się oznaczać, że rozwiązania globalnych problemów wymagają szerokiego zakresu i dużej ilości danych. Oczekuje się, że dzięki technologiom pozwalającym efektywnie wyodrębniać dane i zarządzać nimi, łatwiej będzie znaleźć wyjście z niektórych potencjalnie katastrofalnych sytuacji na świecie.

Oczywiście, gdy po raz pierwszy stykamy się z nazwą Big Data, nasza uwaga skupia się na słowie „Big", a na myśli przychodzi wyobrażenie gigantycznej istoty. W rzeczywistości jednak określenie Big Data jest bardziej związane z ogromem i niepoliczalnością. Pojęcie to zostało zdefiniowane i rozpowszechnione w 2001 r. przez pracującego dla Meta Group (teraz Gartner) analityka Douga Laneya, a odnosiło się ono do problemów i możliwości związanych z trzema wymiarami szybkiej ekspansji danych: ilości, szybkości przepływu i różnorodności [4]. Szersze zainteresowanie pojęciem Big Data w pierwszej dekadzie XXI wieku można powiązać z globalnym rozpowszechnieniem dostępu do Internetu i koniecznością analizowania generowanych przez niego ogromnych ilości danych. Nie sposób przecenić znaczenia analizowania przytłaczających ilości danych i przekształcania ich w użyteczne informacje. A do trzech wspomnianych wymiarów danych należy dodać „wartość". Jeśli potraktujemy Big Data jako ogromne ilości danych generowane w czasie rzeczywistym na podobieństwo strumieniowania, w tym danych niestrukturalnych, takich jak tekst, obrazy i wideo, ważna jest umiejętność połączenia tych różnych typów danych w celu utworzenia wartości. Znaczenie ma ilość zasobów, a nie rozmiar kopalni. Badacz nie potrzebuje danych; tym, czego szuka, są informacje. Big Data odnosi się do rozmiaru danych; w gruncie rzeczy ważniejsza jest jednak analiza i wytwarzanie danych mających sens.

Aby można było mówić o Big Data, zbiór danych musi mieć duży rozmiar. Wprawdzie nie ma żadnego zdefiniowanego limitu, od którego zaczynamy mówić o Big Data, ale typowy zbiór danych Big Data liczy od kilku terabajtów (mały) do kilku petabajtów (duży). W tabeli 1.1 znajdują się informacje o wykorzystywanych obecnie wielkościach danych, które wyraża się za pomocą przyrostków peta-, eksa-, zetta-, jotta-, bronto- i geop- [5]. Gdybyśmy zechcieli w ten sposób wyrazić ilość danych zawartych w książkach znajdujących się w waszyngtońskiej Bibliotece Kongresu, ich suma wyniosłaby ok. 15 TB. Do roku 2012 ludzkość zgromadziła dane w ilości 1,27 ZB. Wynika z tego, że 1 GpB to ilość danych, jaką trudno sobie wyobrazić. Opisuje ona jednak dane tworzone i rozpowszechniane.

Kolejnym aspektem Big Data jest szybkość przepływu i gromadzenia danych. Dwadzieścia lat temu, zarówno instalacja szybkiej sieci przepływu danych, jak i comiesięczne opłaty za połączenie oznaczały duże wydatki. Obecnie natomiast nie ma

Tabela 1.1 Porównanie jednostek danych

Dane	Rozmiar		Przykłady				
Bit (b)	1 b	1	Cyfra binarna (1 lub 0)				Podstawowe jednostki danych
Bajt (B)	8 b	2^3	Angielska litera (1 znak)				
Kilobajt (kB)	1024 B	2^{10}	1	Strona			Arkusz papieru zawierający 1200 znaków
Megabajt (MB)	1024 kB	2^{20}	873	Strony	4	Książki	Jedno cyfrowe zdjęcie: 3 MB Jedna piosenka MP3: 4 MB
Gigabajt (GB)	1024 MB	2^{30}	894 784	Strony	4 473	Książki	Film o długości 12 godzin: 12 GB
			341	Cyfrowe obrazy	256	Pliki dźwiękowe MP3	
Terabajt (TB)	1024 GB	2^{40}	916 259 689	Strony	4 581 298	Książki	Wszystkie książki w Bibliotece Kongresu: 15 TB
			349 525	Cyfrowe obrazy	262 144	Pliki dźwiękowe MP3	
			1 613	Płyty CD	233	DVD	
			40	Dyski Blue-ray			
Petabajt (PB)	1024 TB	2^{50}	938 249 922 368	Strony	4 691 249 611	Książki	Ilość danych przetwarzanych przez Google w każdej godzinie: 1 PB
			357 913 941	Cyfrowe obrazy	268 435 456	Pliki dźwiękowe MP3	
			1 651 910	Płyty CD	239 400	DVD	
			41 943	Dyski Blue-ray			
Eksabajt (EB)	1024 PB	2^{60}	960 767 920 505 705	Strony	4 803 839 602 528	Książki	Ilość danych zawartych w 100 milionach kopii amerykańskiego tygodnika
			366 503 875 925	Cyfrowe obrazy	274 877 906 944	Pliki dźwiękowe MP3	
			1 691 556 350	Płyty CD	245 146 535	DVD	
			42 949 672	Dyski Blue-ray			

Tabela 1.1 c.d.

Dane	Rozmiar		Przykłady				
Zettabajt (ZB)	1024 EB	2^{70}	983 826 350 597 842 752	Strony	4 919 131 752 989 213	Książki	Ilość danych istniejących do roku 2012: 1,27 ZB
			375 299 968 947 541	Cyfrowe obrazy	281 474 976 710 656	Pliki dźwiękowe MP3	
			1 732 153 702 834	Płyty CD	251 030 052 003	DVD	
			43 980 465 111	Dyski Blue-ray			
Jottabajt (YB)	1024 ZB	2^{80}	1 007 438 153 012 190 978 921	Strony	5 037 190 915 060 954 894	Książki	Na pobranie 1 YB danych przez szybkie łącze trzeba by 11 trylionów lat
			3843 307 168 202 282 325	Cyfrowe obrazy	288 230 376 151 711 744	Pliki dźwiękowe MP3	
			1 773 725 391 702 841	Płyty CD	257 054 773 251 740	DVD	
			45 035 996 273 704	Dyski Blue-ray			
Bronto-bajt (BB)	1024 YB	2^{90}	1 031 616 699 404 483 562 415 936	Strony	5 158 083 497 022 417 812 079	Książki	Biorąc pod uwagę dane, które można zebrać w czasie rzeczywistym za pomocą czujników IoT (Internet of Things)
			393 530 540 239 137 101 141	Cyfrowe obrazy	295 147 905 179 352 825 856	Pliki dźwiękowe MP3	
			1 816 294 801 103 709 697	Płyty CD	263 224 087 809 782 414	DVD	
			46 116 860 184 273 879	Dyski Blue-ray			
Geopbajt (GpB)	1024 BB	2^{100}	1 056 375 500 190 191 167 913 919 337	Strony	5 281 877 500 950 955 839 569 596	Książki	Największa ilość danych, jaką można sobie wyobrazić
			402 975 273 204 876 391 568 725	Cyfrowe obrazy	302 231 454 903 657 293 676 544	Pliki dźwiękowe MP3	
			1 859 885 876 330 198 730 317	Płyty CD	269 541 465 917 217 192 562	DVD	
			47 223 664 828 696 452 136	Dyski Blue-ray			

zwykle problemu z używaniem połączeń sieci przewodowych i bezprzewodowych do transferu 1 Gb/s (może to być, przynajmniej w Korei, 100 Mb/s) z domu, z biura, a nawet z ulicy. Dzięki temu tworzenie i dystrybucja danych odbywają się w mgnieniu oka. W ostatnich czasach informacje o katastrofach naturalnych i innych ważnych zdarzeniach jako pierwsze podawały nie telewizyjne wiadomości, ale mikroblogi takie jak Twitter. Co więcej, inteligentne mierniki w zakładach przemysłowych, urządzeniach AGD takich jak inteligentne telewizory i lodówki, oraz samochody bez kierowcy coraz częściej są podłączone do Internetu, co umożliwia gromadzenie w czasie rzeczywistym ogromnych ilości danych, których będzie coraz więcej.

O Big Data należy mówić nie tylko w kategoriach rozmiaru i szybkości wynikających z nieustannego gromadzenia różnorodnych danych. W przeszłości wszystkie dane, które wykorzystywaliśmy w pracy, były dobrze sformatowane i łatwo było nimi zarządzać. Innymi słowy, dane były uporządkowane i miały konkretną postać, a więc miały strukturę. Do takich typowych danych można na przykład zaliczyć wyniki sprzedaży, dane magazynowe albo wskaźniki awaryjności w czasie przetwarzania. Natomiast nowe typy danych nie mieszczą się w dotychczasowych schematach, ponieważ są niestrukturalne. Wideo, muzyka, obrazy, informacje o lokalizacji, tekst itp. to dane, które nie pasują do istniejących formatów; są to dane niestrukturalne. Mają one różne rozmiary i zawartość, którą trudno uporządkować, ale jest ich znacznie więcej; dlatego też uzyskiwanie danych, które mają sensowne znaczenie, wymaga zastosowania nowych metod przetwarzania.

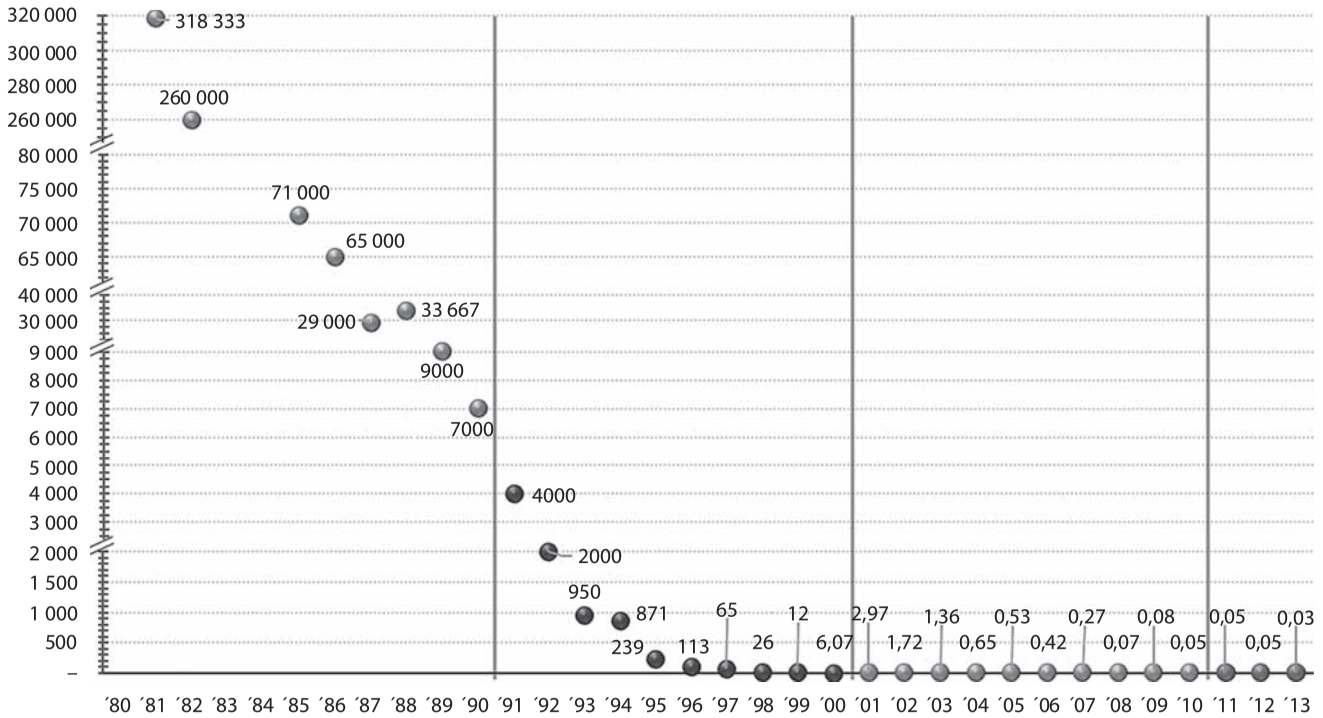
Biorąc pod uwagę rozmiar, szybkość przepływu i różnorodność, typowy zbiór Big Data ma około kilku tysięcy terabajtów, wytwarzanych, rozprowadzanych i wykorzystywanych w czasie od kilku sekund do paru godzin, a zebrane dane mogą mieć postać strukturalną lub niestrukturalną, co sprawia, że analiza tych danych i zarządzanie nimi za pomocą dotychczasowych metod są w praktyce niemożliwe. Ponadto, mimo ogromnych ilości i szybkości tak różnorodnych postaci danych, w gruncie rzeczy najważniejsze jest rozpoznawanie tylko istotnych danych poprzez analizę Big Data. Dlatego też Big Data, oprócz zbiorów danych, którymi trudno zarządzać i które trudno analizować tradycyjnymi metodami, to również zasoby ludzkie, organizacje i odpowiednie technologie do analizowania złożonych danych i zarządzania nimi. I właśnie w tym procesie wytwarzana jest wspomniana wartość Big Data.

We współczesnym, połączonym społeczeństwie, ilość danych jest ogromna. Jednak sam ten fakt nie oznacza jeszcze, że trzeba je wszystkie analizować. Wśród rosnącej ilości danych wzrasta również ilość danych bez znaczenia, kompletnie nieprzydatnych, a uzyskanie na ich podstawie zbiorów danych mających znaczenie, wymaga specjalnych umiejętności. W analizie Big Data kładzie się nacisk na rozszerzanie perspektyw i horyzontów myślowych; wiąże się to nie tylko z ilością, szybkością, czy rozmiarem danych, ale również z umiejętnością znalezienia przemyślanego punktu widzenia

i prognozowania. Aby zobaczyć las, nie trzeba z niego wychodzić; wystarczy wejść na szczyt góry. Podobnie rzecz ma się z Big Data: musimy podnieść perspektywę na wyższy poziom. Z góry widok jest rozleglejszy niż z poziomu ziemi. Znajdując się wewnątrz grupy nie można przyjrzeć się jej z boku, konieczna jest inna perspektywa. Aby zobaczyć jeszcze więcej, musimy sięgnąć po Big Data. Analiza Big Data pozwala nam znacząco poszerzyć dotychczasowe pole widzenia.

1.2. Co tworzy Big Data?

W całej historii (do roku 2012) ludzkość zgromadziła około 1,27 ZB danych i szacuje się, że do roku 2016 globalny ruch IP (Internet Protocol) osiągnie w przybliżeniu 1,3 ZB [6]. Nie sposób w pełni pojąć przyczyny tego ogromnego napływu danych. Jedną z nich jest rozwój urządzeń pamięci. Wynalezienie pisma, a potem papieru i druku, poprawiło trwałość zapisów historycznych, do których wcześniej wykorzystywano m.in. łupiny roślin, zwierzęce skóry, zdrewniałe fragmenty roślin (np. bambusowe listewki), kamienne, a później gliniane tabliczki. Drukując znaki na papierze, zaczęto zachowywać wiedzę o wielu dziedzinach działalności człowieka, która kiedyś znikała bez śladu. Jednak cechą informacji zapisanych na papierze jest objętość przerastająca ich ilość. W XX wieku pojawiły się analogowe urządzenia pamięci takie jak błony fotograficzne, płyty dźwiękowe, kasety magnetofonowe i kasety wideo, a informacje zapisane dawniej na papierze były teraz składowane na kliszach lub taśmach o mniejszej objętości, które mogły zawierać znaczące ilości danych. W latach 80. XX w., przed nadejściem ery cyfrowej, ludzkość zgromadziła około 2 620 000 TB danych, z czego co najmniej 90% znajdowało się na filmach i taśmach [7]. Po roku 1990 nadeszła rewolucja cyfrowa, czyli cyfryzacja znaków, dźwięków, obrazków i obrazów, z którą związany był radykalny wzrost możliwości przechowywania danych. Pierwszym komputerowym urządzeniem pamięci była dyskietka elastyczna; później pojawiły się takie urządzenia jak dyski twarde i pamięć flash. Dzisiaj, na naszych smartfonach możemy codziennie zapisywać i przeglądać dziesiątki GB, w dowolnej chwili pobierając wedle uznania książki, obrazy, muzykę i zdjęcia. Gdyby w tej chwili dane zgromadzone przez ludzkość zapisać na płytach CD i ułożyć je w stos, powstałaby wieża o wysokości sześciokrotnie większej niż odległość z Ziemi do Księżyca. Dzięki osiągnięciom technologicznym wciąż spada cena pamięci. Zapisanie 1 GB danych na dysku twardym w 1980 r. kosztowało 213 tysięcy USD. W roku 2013 zapisanie tej samej ilości danych na dysku twardym kosztowało 3 centy [8]. Ten spadek kosztów ma istotny wpływ na gwałtowny wzrost gromadzenia danych (patrz rysunek 1.1).



Rysunek 1.1. Średni koszt dysku twardego na gigabajt (jednostka: USD)

Za kolejną przyczynę istotnego wzrostu ilości danych można uznać coraz lepszą łączność. W latach 60. XX w. komputery były rzadkie i drogie, ale wraz z pojawieniem się komputerów osobistych w latach 80., ich wynikające z wyjątkowości ceny zaczęły spadać. Dzisiaj obserwujemy coraz większą dostępność przenośnych komputerów i urządzeń osobistych, w tym inteligentnych urządzeń komunikacyjnych takich jak smartfony i cyfrowe tablety połączone z Internetem, co sprawia, że wartość wynikająca z wyjątkowości komputerów zanika. Wiele posiadanych dziś powszechnie smartfonów ma wydajność o wiele wyższą od niektórych wyprodukowanych zaledwie kilka lat temu komputerów osobistych, dzięki czemu można używać ich do sterowania różnymi urządzeniami. A od czasu pojawienia się na rynku inteligentnych telewizorów i lodówek, rozwija się i upowszechnia łączność komputerów z różnymi urządzeniami mającymi zdolność do komunikacji bezprzewodowej – od pojazdów po urządzenia AGD. W efekcie narodziło się pojęcie IoT (ang. *Internet of Things*, Internet rzeczy), które dotyka strukturalnego aspektu technologii łączności internetowej, polegającego na gromadzeniu danych z czujników zainstalowanych w różnych obiektach. Według danych firmy Gartner, w 2009 r. z technologii IoT korzystało 900 milionów jednostek; oczekuje się, że do roku 2020 ta liczba urośnie do 26 miliardów. Cisco spekuluje, że wartość ekonomiczna technologii IoT w latach 2013–2022 wyniesie w przybliżeniu 14 bilionów USD.

Siłą napędową znaczenia Big Data nie może być jedynie wykładniczy przyrost danych. Cały czas rośnie ilość i szybkość różnego typu danych, więc kluczowe znaczenie ma wyłuskiwanie z Big Data przydatnych informacji. Aby wyodrębnić przydatne informacje, potrzebne są techniki zarządzania i analizy danych. Przed ostatnią dekadą XX wieku, przeciętna baza danych cyfrowych obrazów zawierała dziesiątki tysięcy obrazów. Natomiast dzisiejsze witryny do udostępniania zdjęć, takie jak Flicker, Picasa i Pinterest, mają bazy danych cyfrowych obrazów o rozmiarach przekraczających naszą wyobraźnię; każda z nich może zawierać więcej niż dziesiątki miliardów obrazów. Gdyby nie ewolucja technologii przetwarzania danych graficznych, to nawet przy rosnącej wydajności komputerów, niemożliwe byłoby zarządzanie tak szybko rosnącą ilością danych. Na szczęście tempo rozwoju technologii przetwarzania obrazów cyfrowych, polegającej na analizowaniu i indeksowaniu obrazów, dorównywało szybkości wzrostu ilości danych. Dzięki temu możliwe stało się przeszukiwanie miliardów obrazów na sekundę, a to z kolei pozwoliło na zarządzanie ogromnymi ilościami danych. Analogiczne osiągnięcia pojawiły się w analizie różnego typu danych niestrukturalnych i związanych z tym technologiach. Pozwoliły one docenić wartość Big Data, i stanowią obecnie niezwykle popularne zagadnienie.

Oczywiście, osiągnięcia w świecie technologicznym nie zawsze przyciągają uwagę rynku. Często technologia, która jest dostępna od dłuższego czasu, staje się popularna, dopiero gdy powiąże się ją z potrzebami rynku. Wielu pionierskim w swoim

czasie technologiom wróżono świetlaną przyszłość, ale słuch o nich zaginął, zanim jeszcze zaistniały na rynku, ponieważ konsumentom nie były do niczego potrzebne. Pod tym względem pytanie, jak stosować Big Data w działalności korporacji staje się niesłychanie ważne.

1.3. Jak używamy Big Data?

W jaki sposób możemy, korzystając z Big Data, udoskonalić środowisko biznesowe? Zmiana naszej perspektywy to pierwszy krok na drodze do jego ulepszenia. Analiza Big Data i oddzielenie „ziarna od plew" nie polega na ocenie tego, co już wiemy, lecz raczej odkryciu tego, czego nie wiemy. Ocena to weryfikacja naszej wiedzy, natomiast odkrywanie to ustalanie, jakie pytania należy postawić w trakcie kreatywnego, wielokrotnie powtarzanego procesu eksploracji. Odkrywanie to tworzenie prawdziwej wartości z Big Data. W ten sposób firma zyskuje pomysł na zwiększenie swojej wartości korporacyjnej [9].

Drugi krok na drodze do ulepszenia korporacyjnego środowiska biznesowego polega na odkryciu różnych problemów i możliwych rozwiązań w różnych aspektach działalności firmy. Obejmuje to zmiany w korporacyjnym procesie myślowym i metodach podejmowania decyzji za pomocą analizy Big Data. Analiza ta umożliwia znalezienie ważnej, ukrytej, nieznanej wcześniej prawdy. Znalezienie w Big Data istotnej prawdy, do odkrycia której nie wystarczają ludzkie zdolności poznawcze, to proces rozwiązywania problemów za pomocą analizy Big Data. Prognozowanie to kolejne działanie realizowane w ramach tej analizy. Przyszłość nie jest oddzielona od teraźniejszości ani od przeszłości. Wybory dokonane w przeszłości istnieją w teraźniejszości, a to, co wybieramy w teraźniejszości przechodzi do przyszłości. Ponieważ przeszłość, teraźniejszość i przyszłość nie są oddzielone, to poznając teraźniejszość, można otworzyć drzwi do przyszłości. Analiza różnych przeszłych postaci Big Data może znacznie ułatwić ujrzenie możliwych rezultatów i przyszłych okoliczności. Analizując ogromne ilości danych, ludzie mogą odkrywać nowe obszary wiedzy i zdarzenia, a dzięki tym odkryciom łączyć teraźniejszość z przyszłością. Innymi słowy, prawidłowa analiza danych umożliwia prognozowanie. Klasycznym przykładem jest marketingowa analiza koszyków sklepowych, czyli badanie zachowania klientów w związku z dokonywaniem wyborów przy zakupach; obejmuje ona analizę *cross-selling* (sprzedaż krzyżowa) i *up-selling* (sprzedaż dodatkowa), czyli sprzedaż optymalizowaną poprzez selektywne rozmieszczanie towarów na sklepowych półkach. Analiza zachowania klientów dokonujących zakupów wykazała, że klienci, którzy kupują pieluchy, mają również skłonność do kupowania piwa, a przed nadejściem huraganu kupowano najczęściej

latarki i słodkie przekąski. Na podstawie tych ustaleń rozmieszcza się produkty na sklepowych półkach, aby klienci kupowali jeszcze więcej: piwo jest ustawiane w pobliżu pieluch, a ciastka truskawkowe koło latarek, czego skutkiem jest zwiększenie sprzedaży. Ważnym i coraz częściej podkreślanym elementem analizy Big Data jest wizualizacja wyników. Wizualizacja to techniki i metody ułatwiające zrozumienie danych „na pierwszy rzut oka". Pracownicy korporacji muszą jasno rozumieć, co zostało wykryte poprzez analizę Big Data; to wymaganie sprawia, że wizualizacja jest niezwykle przydatna.

Dane istnieją w każdej korporacji, a ich dostępność umożliwia zwiększenie wydajności. Trzecim krokiem na drodze do ulepszenia korporacyjnego środowiska biznesowego, polegającym na zastosowaniu różnych punktów widzenia na istniejące dane, jest usprawnienie procedur wykorzystywania istniejących danych i technologii informacyjnej w celu podniesienia wydajności pracy. O zwiększaniu wydajności mówi się często i choć brzmi to jak banał, w środowisku korporacyjnym jest to cały czas istotna kwestia. Metody używania Big Data do podnoszenia wydajności korporacyjnej można podzielić na dwie kategorie. Po pierwsze, korzystając z czujników, można gromadzić dane na temat ruchu materiałów i zarządzać nimi, co zmniejsza koszty pracy, magazynowania i logistyki. Po drugie, poprzez analizę Big Data można zminimalizować zbędne działania w przepływie pracy łańcucha wartości, co pozwala na restrukturyzację procesu pracy i dzięki temu – optymalizację wydajności. Do tej pory podnoszenie wydajności w przemyśle osiągano, zastępując ludzką siłę roboczą maszynami i komputerami; natomiast w świecie Big Data surowce, produkty, maszyny itd. są wyposażane w różne czujniki i znaczniki, umożliwiające interakcję w czasie rzeczywistym i akumulację danych w celu zwiększenia wydajności. W porównaniu z przeszłością, dane są bardziej miarodajne, szybciej generowane i akumulowane, można więc mówić o innym wymiarze zwiększania wydajności. Za sprawą danych generowanych i rejestrowanych przez różne czujniki w czasie rzeczywistym, możliwe jest wykonywanie bardziej szczegółowych prognoz, co pozwala na bardziej precyzyjne zarządzanie. Na przykład w SCM (*Supply Chain Management* – zarządzanie łańcuchem dostawczym) występuje tzw. efekt bykowca (ang. *bullwhip effect*). Jest to błąd wynikający z nieprawidłowego oszacowania dostawy surowców, który zwiększa się w trakcie prognozowania wyników produkcji w kolejnym ogniwie łańcucha dostaw, powodując większą rozbieżność w porównaniu z rzeczywistymi wynikami sprzedaży. Wynika z tego, że dzięki precyzyjnym prognozom i kontrolom przeprowadzanym na podstawie danych w czasie rzeczywistym, można zminimalizować rozbieżność pomiędzy podażą a popytem, a tym samym obniżyć koszty magazynowania i logistyki, co pozwala zwiększyć wydajność. I faktycznie, system zarządzania sieciami elektroenergetycznymi następnej generacji, Smart Grid, który jest przykładem powiązania technologii informacyjnej (IT) z tradycyjnymi systemami energetycznymi, opiera się

na instalowaniu różnych czujników i mierników dostarczających w czasie rzeczywistym informacje na temat trendów zużycia prądu, dzięki którym można uzyskać większą wydajność i skutecznie zarządzać produkcją mocy. Ta dwustronna komunikacja umożliwia szczegółową kontrolę i zdalne sprawdzanie urządzeń, którymi płynie prąd. Możliwe jest również automatyczne przywracanie normalnego funkcjonowania w wypadku zakłóceń (przerw), a także optymalizacja dystrybucji mocy zgodnie z faktycznym zużyciem i szybsze reakcje w przypadku sytuacji awaryjnych. W Stanach Zjednoczonych, firmy zajmujące się dostarczaniem surowców i usług dla produkcji odchodzą od tradycyjnych metod przeprowadzania napraw po awarii sprzętu na rzecz „monitoringu na podstawie stanu", który polega na aktywnym śledzeniu stanu urządzeń, zanim nastąpi awaria. Czekanie do chwili wystąpienia awarii powoduje nieprzewidziany spadek wydajności związany z przerwą w pracy linii produkcyjnej podczas jej naprawy. Śledząc na bieżąco np. temperaturę, wibracje, rozmiar produkcji i wysyłając te dane do analizy przez system, inżynierowie mogą zawczasu przygotować się na wystąpienie problemu i opracować rozwiązania, które nie będą miały wielkiego wpływu na proces produkcji. Do podniesienia wydajności przyczynić się mogą nie tylko dane strukturalne, lecz również niestrukturalne. Można np. przeanalizować procesy tworzenia i dystrybucji dokumentów w firmie i je udoskonalić, aby zwiększyć wydajność. Usprawnienie takich działań jak wyszukiwanie dokumentów lub obrazów może spowodować zmniejszenie kosztów i czasu uzyskiwania informacji, i jest to kolejny obszar zastosowania Big Data.

Czwartym krokiem na drodze do ulepszenia korporacyjnego środowiska zarządzania jest zapewnienie obiektywnego punktu widzenia osobom odpowiedzialnym za podejmowanie decyzji. Każde nowe wyzwanie wymaga podjęcia kilku decyzji i przy tej okazji pojawia się wiele konfliktów. Jeśli dostępne są jakieś obiektywne, przez nikogo niekwestionowane dane, decydentowi łatwiej przezwyciężyć uprzedzenia i pokonać słabości, czego skutkiem może być racjonalny kompromis w ramach organizacji. Podejmowanie decyzji zawsze będzie podstawowym obowiązkiem menedżera, a niepowodzenie wielu korporacji jest wynikiem podjęcia złych decyzji. Intuicja doświadczonej osoby, zwykle bardzo ważna, czasami może stanowić przeszkodę przy podejmowaniu racjonalnych decyzji. Gdy intuicję wesprze się wynikami analizy Big Data, można wyciągać trafniejsze wnioski. Na przykład w branży paliwowej, poszukiwanie i planowanie pól naftowych i gazowych wykonuje się, instalując ogromne sieci sensorów w skorupie ziemskiej, by precyzyjnie określić rozmieszczenie i strukturę pól naftowych. Efektem tego są niższe koszty budowy infrastruktury i transportu ropy. Aby efektywnie wykorzystywać Big Data w korporacyjnym procesie podejmowania decyzji, konieczne jest właściwe pokierowanie przepływem przetwarzania danych. Ponieważ same ogromne zbiory danych nie biorą udziału w procesie podejmowania decyzji, w korporacji należy wdrożyć optymalny przepływ tego procesu. Ostatni krok na drodze do ulepszenia korporacyjnego środowiska zarządzania polega

na utworzeniu nowej wartości i zintegrowaniu jej z nowym planem biznesowym. Rozsądne wykorzystanie danych pozwoli stworzyć nowy paradygmat zarządzania biznesowego. Ostatecznym celem stosowania analizy Big Data jest znalezienie tego, co przeoczono w przeszłości, jak również odkrycie ukrytej wartości dynamicznego klienta albo utworzenie nowej wartości i dostarczenie jej klientowi. W przemyśle energetycznym do zmiany wzorców zużycia energii wykorzystuje się połączone z siecią czujniki i automatyczne mechanizmy dostarczania informacji zwrotnej. Instalując inteligentne mierniki w przemysłowych elektrowniach, można weryfikować koszty zużycia energii w czasie rzeczywistym i na tej podstawie ustalić okresy szczytowego użycia energii w godzinach roboczych, a następnie zapewnić optymalne sterowanie przeciążeniem sieci energetycznej. To z kolei pozwala na przeniesienie energochłonnych procesów na czas mniejszego obciążenia sieci. Przykładem może być jedno z ostatnich osiągnięć w systemach nawigacji samochodowej. Systemy te to dzisiaj więcej niż tylko połączenie elektronicznej mapy i systemu GPS (*global positioning system*) z informacjami zebranymi z czujników drogowych, które są przesyłane do systemu nawigacji, a następnie na centralny serwer. Po analizie danych pod kątem warunków ruchu klient otrzymuje informację o optymalnej, najkrótszej drodze do wybranego celu. Dzięki tym technologicznym innowacjom i analizie Big Data spada zużycie paliwa, a klienci, unikając stania w korkach, mogą efektywniej wykorzystać swój czas. Wydaje się, że w najbliższej przyszłości można będzie opracować system nawigacji, który na podstawie harmonogramu zajęć i nastroju kierowcy będzie potrafił wskazać optymalne kierunki jazdy, jeszcze zanim kierowca ich zażąda. Pamiętając o tle tych ewolucji, nie można przecenić znaczenia zbierania i analizy szczegółowych danych.

Aby jednak efekt stosowania analizy Big Data był jak największy, wymagana jest technologia efektywnego przetwarzania danych. Najprościej rzecz ujmując, chodzi o odpowiedź na pytanie, jaki problem powinien przejść określony proces i kto będzie zarządzać tym procesem, aby uzyskać największą wydajność. Zwykle nie koncentrujemy się na samym problemie, ale na tym, jak szybko można go rozwiązać. Dlatego też najważniejszym aspektem jest zidentyfikowanie właściwego problemu. Przy zbieraniu informacji na temat problemu przydatna może być analiza Big Data. Ogólnie rzecz biorąc, odbywa się ona w trzech etapach. Pierwszy etap to obserwacja. Analiza Big Data nie wymaga zbierania dużych ilości danych. Dane stają się istotne dopiero podczas analizy. Dlatego na początku musimy na podstawie obserwacji zdecydować, jakie dane chcemy zbierać. Następnie konieczne jest ujęcie ilościowe. Jest to proces niezbędny do systemowej obserwacji wykraczającej poza proste zbieranie danych. W analizie Big Data wymagana jest szersza, elastyczna metoda analizy ilościowej. Wreszcie potrzebny jest proces rozumowania dedukcyjnego. Na przykład ilość danych znacznie wzrosła wraz z rozpowszechnieniem się smartfonów. Stosując dedukcję musimy zastanowić się nad tym, dlaczego ludzie używają smartfonów.