PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 508 • 2018

Taksonomia 31
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

Kamila Migdał-Najman, Krzysztof Najman
Uniwersytet Gdański
e-mails: kamila.migdal-najman@ug.edu.pl; krzysztof.najman@ug.edu.pl

PROFILOWANIE, OCZYSZCZANIE I ZAPOBIEGANIE POWSTAWANIU *DIRTY DATA*

DIRTY DATA – PROFILING, CLEANSING AND PREVENTION

DOI: 10.15611/pn.2018.508.15
JEL Classification: C38, C55, C81, C82

Streszczenie: Zbiory Big Data oferują dostęp do niemal nieograniczonej liczby danych, dając nadzieję na szybszy, tańszy, bardziej precyzyjny i wszechstronny opis świata. Jednocześnie w takich zbiorach poza danymi o odpowiedniej jakości (*clear data*) znaczny udział mają dane nieprawdziwe, nieaktualne, zaszumione, często zwielokrotnione, niepełne lub błędne (*dirty data*), a także dane o nieznanej jakości czy użyteczności (*dark data*). Znaczący udział *dirty* i *dark data* ma szereg negatywnych konsekwencji w analizie zbioru Big Data. Celem prezentowanych badań jest przegląd i systemowe ujęcie procedur minimalizowania negatywnych efektów *dirty data* w analizie Big Data. W konstrukcji systemu oczyszczania zbioru danych uwzględniono najważniejsze procedury profilowania (*profiling data*), oczyszczania (*cleansing data*) i zapobiegania (*defect prevention*) powstawaniu *dirty data* w procesie budowy i analizy zbioru Big Data.

Słowa kluczowe: big data, dirty data, profilowanie danych, oczyszczanie danych, zapobieganie powstawaniu zanieczyszczeń w danych.

Summary: There are almost unlimited sources of large streams of information now being referred to as Big Data. Because of it we hope for a faster, cheaper, more precise and versatile description in the world around us. At the same time, in such data sets, apart from data of a proper quality (clear data), significant share is false, outdated, noisy data, often multiplied, incomplete or incorrect (dirty data), as well as data of unknown quality or usefulness (dark data). A significant share of dirty data and dark data causes a number of negative consequences in the analysis of Big Data sets. The aim of this article is to review and systemically capture the procedures for minimizing the negative effects of dirty data in the analysis of Big Data. The design of the data collection system includes the most important profiling procedures (profiling data), cleansing data and defect prevention of dirty data in the process of building and analyzing the Big Data sets.

Keywords: Big Data, dirty data, profiling data, data cleansing, defect prevention.

1. Wstęp

W roku 2017 mija równo 20 lat od momentu, w którym pracownicy NASA zgłosili swoim pracodawcom szczególny problem. Mimo dostępu do superkomputerów, ogromnej infrastruktury IT, potężnych środków finansowych nie byli w stanie przeprowadzić wizualizacji wyników swoich badań [Cox, Ellsworth 1997][1]. Na pytanie przełożonych, co jest przyczyną ich problemów, padła odpowiedź, że opracowywane dane nie mieszczą się na dyskach lokalnych NASA, dyskach zdalnych, z których mogą korzystać, nie mieszczą się także w pamięci RAM komputerów. Mają problem z dużą ilością danych. Po raz pierwszy w literaturze pojawiło się pojęcie Big Data [Cox, Ellsworth 1997; Zomaya, Sakr (red.) 2017].

Współcześnie ze zjawiskiem tego typu mamy do czynienia w coraz większej liczbie organizacji. Rozwój techniki teleinformacyjnej, Internetu i informatyki przy jednoczesnym spadku jednostkowych kosztów gromadzenia i przechowywania danych powoduje, że możliwe staje się zbieranie praktycznie dowolnej ilości[2] danych. Pochodzą one z wielu źródeł. Najważniejsze, z biznesowego punktu widzenia, są obecnie dane będące efektem interakcji międzyludzkich (*human interaction data*). Tworzą je wszelkie formy komunikacji, takie jak wiadomości e-mail, sms, wszelkie przesyłane dokumenty tekstowe, zdjęcia, filmy czy nagrania dźwiękowe [Migdał-Najman, Najman 2017]. Stanowią one podstawę wielu modeli biznesowych związanych z promocją, reklamą, sprzedażą i dystrybucją niezliczonych produktów i usług. Aby te działania mogły być skuteczne, adekwatne modele muszą opierać się na danych o odpowiedniej jakości. Niestety zbiór Big Data jest zbiorem rozproszonym, niestrukturyzowanym, powstającym jednocześnie w ogromnej liczbie miejsc, w różnych systemach czy standardach. Powoduje to, że obok danych wysokiej jakości (*clear data*) znajdują się tu także dane nieprawdziwe, nieaktualne, zaszumione, często wielokrotnie zduplikowane, niekompletne lub błędne (*dirty data*), a także dane, o których jakości czy użyteczności nic nie wiadomo (*dark data*). Znaczący udział *dirty* i *dark* data powoduje szereg negatywnych konsekwencji w analizie Big Data [Migdał-Najman, Najman 2013].

Celem prezentowanych badań jest przegląd i systemowe ujęcie procedur minimalizowania negatywnych efektów *dirty data* w analizie Big Data. W konstrukcji systemu oczyszczania zbioru danych uwzględniono najważniejsze procedury profilowania (*profiling data*), oczyszczania (*cleansing data*) i zapobiegania (*defect prevention*) powstawaniu *dirty data* w procesie budowy i analizy zbioru Big Data.

---

[1] „We call this the problem of big data."

[2] Liczba gromadzonych danych jest tak wielka, że rozważane są systemy o praktycznie nieskończonym wolumenie danych. Zbiór staje się faktycznie niepoliczalny.

2. *Dirty data*

Media społecznościowe, z portalem Facebook na pierwszym miejscu, są największym źródłem danych o internautach. Jest to obecnie najczęściej wykorzystywane w praktyce źródło danych Big Data[3]. Wielu badaczy skupia się na „tagach", „hashtagach", „lajkach", „komciach", „szerach", słowach czy emocjach, które charakteryzują strony i wypowiedzi użytkowników. Według analiz Networked Insights [Luebbe 2015] ogromna część takich danych jest bezwartościowa, ponieważ wcale nie pochodzi od realnych użytkowników. Dane te są generowane przez programy komputerowe (boty) podszywające się pod realnych użytkowników (aż 53%), osoby opłacane przez konkurencyjne firmy (23%) lub są efektem działania spamerów czy celebrytów.

Nie wszystkie „brudne dane" powstają jako celowy szum czy dezinformacja. Dane istniejące w domenie publicznej często są po prostu nieaktualne, niezgodne ze stanem faktycznym lub zapisane w wielu różnych standardach[4]. W USA nawet 20% obywateli zmienia co roku adres zamieszkania, 18% obywateli zmienia numer telefonu komórkowego (niektóre z nich otrzymują nowi abonenci), zmienia się 25-33% adresów e-mail (stare przestają być używane, nie są jednak kasowane). W ciągu jednej godziny zmienia się około 58 adresów istniejących przedsiębiorstw, 11 firm zmienia swoją nazwę, powstaje 41 nowych firm [Bjornaas 2015]. Znaczna część tych informacji nie jest poprawiana na wszystkich stronach, na których się pojawiły. Po kilku latach liczba błędnych czy nieaktualnych informacji sięga milionów wpisów. Na jednych stronach WWW zapisane są „stare" informacje, a na innych „nowe". Jedne i drugie dotyczą pojedynczej osoby czy firmy, wprowadzając do automatycznych analiz „zanieczyszczenia" informacyjne.

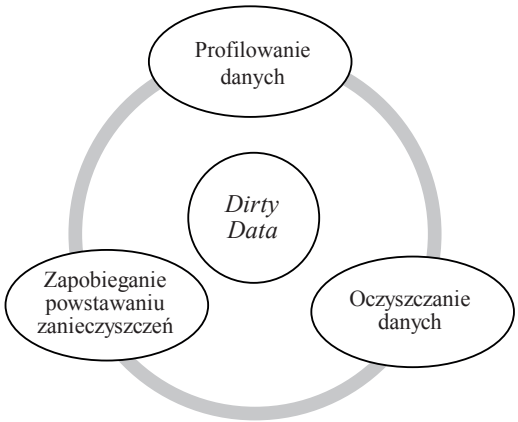
Zanieczyszczenia te mogą mieć różny charakter. Można wyróżnić kilka podstawowych źródeł *dirty data*: 1. Powtórzenia, 2. Niedokładny zapis danych, 3. Błędny zapis danych, 4. Dane niepełne lub niekompletne, 5. Brak integracji informacji pochodzących z różnych źródeł, 6. Informacje naruszające elementarne zasady biznesowe.

Wszystkie te elementy zanieczyszczają zbiór danych, tworząc *dirty data*. Konsekwencje tego zanieczyszczenia są równie wielkie jak samo Big Data. Szacuje się, że w samych Stanach Zjednoczonych problem jakości danych, w tym *dirty data*, generuje rocznie nawet 600 miliardów dolarów kosztów [ReachForce 2015]. Walka z *dirty data* to złożony proces obejmujący analizę profilowania danych (*data profiling*), oczyszczania danych (*data cleansing*) i zapobiegania powstawaniu zanieczyszczeń (*defect prevention*) (rys. 1).

---

[3] Wraz z upowszechnieniem się Internetu rzeczy (*Internet of Things*) proporcje te być może się zmienią.

[4] Typowym przykładem jest zapisywanie imion i nazwisk. Dla automatycznych systemów analitycznych dużym wyzwaniem jest ustalenie, czy pani Yoo Lee ma na imię Yoo, czy Lee?



Rys. 1. Proces przekształcania *dirty data* w *clear data*

Źródło: opracowanie własne.

3. Profilowanie danych

Pierwszym etapem procesu przetwarzania *dirty data* w *clear data* jest **profilowanie danych** (*data profiling*). Jest to statystyczny proces analizy pod kątem ich poprawności, kompletności, unikatowości, spójności, racjonalności i logiki. Aby możliwa była jakakolwiek korekta czy uzupełnienie danych, najpierw trzeba je zrozumieć. Proces profilowania zawiera w sobie analizę kompletności, unikatowości, rozkładu, zakresu, wzorców i powiązań (rys. 2) (szerszą, zorientowaną na realizację informatyczną klasyfikację zob. [Abedjan, Golab, Naumann 2015]).

Rys. 2. Elementy procesu profilowania danych

Źródło: opracowanie własne.

Analiza kompletności (*completeness analysis*) bada, jak często i u których jednostek dana zmienna jest wypełniona, jak często jest pominięta lub jest zerowa. Analiza ta pozwala stwierdzić, czy wypełnienie danego pola jest typowe dla użytkownika, czy nie. W dalszym etapie pozwala weryfikować problemy w zbiorze da-

nych typu: brak wartości zmiennej, a powinna być, jest wartość zmiennej, której nie powinno być.

Analiza unikatowości (*uniqueness analysis*) bada, ile różnych, unikatowych wartości przyjmuje dana zmienna w zbiorze danych. Analiza ta pozwala stwierdzić, czy pojawiają się w zbiorze duplikaty wartości. Pozwala także podjąć decyzję, czy duplikaty są dopuszczalne, czy też nie[5]. W dalszej analizie pozwala weryfikować problemy w zbiorze danych typu: dana wartość jest powtórzeniem innej, a nie powinna być. Pozwala także identyfikować te same jednostki przy łączeniu danych pochodzących z różnych źródeł.

Analiza rozkładu wartości (*values distribution analysis*) bada, jaki jest rozkład częstości danej zmiennej. Jedną ze stosowanych technik jest analiza Benforda (*Benford's Law*) [Benford 1938]. W dalszej analizie pozwala stwierdzić, czy dana wartość pojawia się zbyt rzadko, często czy odpowiednio często w zbiorze danych[6].

Analiza zakresu (*range analysis*) bada, jakie są wartości ekstremalne (a także typowe) danej zmiennej. W dalszej analizie pozwala identyfikować błędy przekroczenia możliwych zakresów zmiennych[7]. Pozwala także identyfikować zmienne, dla których wartości typowe odbiegają od znanych lub możliwych do oszacowania wielkości. Pozwala to na wykrywanie wartości skrajnych (*outliers*), które mieszczą się w zakresie teoretycznym, jednak w znaczącym stopniu wpływają na wartość parametrów rozkładu.

Analiza wzorców (*pattern analysis*) bada, czy dane odpowiadają przyjętemu formatowi lub wzorcowi kodowania. Identyfikuje błędy niezgodności typów takich jak np. ciągły – skokowy (zapisana wartość jest liczbą rzeczywistą, a powinna być liczbą naturalną), tekstowy – liczbowy (zapisana wartość jest typu tekstowego, a powinna być liczbą), liczbowy – data (zapisane liczby nie są datą), liczbowy – kod pocztowy (zapisane liczby nie tworzą kodu pocztowego). Dla niektórych typów danych w dalszej analizie pozwala określić właściwy format danych i metody poprawiania błędnych wpisów. Typowym przykładem jest zapis numerów telefonów, np. (00) 123-456-789, (00) 12-34-56-789, (00) 123-45-67-89 czy 00123456789. Analiza wzorców pozwala rozpoznać, czy kod jest kompletny, możliwy do poprawienia czy uzupełnienia.

Analiza powiązań (*dependency analysis*) bada zależności występujących dla danej zmiennej lub między wieloma zmiennymi w zbiorze danych. Zależności te mogą być wielorakiego rodzaju: 1. Logiczne – jeżeli wiadomo, że elementy typu

---

[5] Typowym przykładem są różnego rodzaju identyfikatory, które z definicji powinny być jednoznaczne, a więc także unikatowe, niepowtarzalne. O ile kilka osób może mieć to samo imię i nazwisko, o tyle PESEL powinien już być identyfikatorem unikatowym.

[6] Jeżeli badamy populację, o której wiemy, że w przybliżeniu zawiera 50% kobiet i 50% mężczyzn, to w analizie zmienna płeć powinna mieć takie właśnie proporcje. Automatyczne rozpoznawanie płci na podstawie informacji z wielokulturowych i wielojęzycznych mediów jest zadaniem trudnym.

[7] Nikt nie może być młodszy niż zero lat, nie może dojeżdżać codziennie do pracy dłużej niż 12 godzin, nie może być w pracy dłużej niż 24 godziny na dobę itp.

A zawierają się w B, to wiadomo, że wszystkie podtypy A1, A2, A3, … także zawierają się w B. 2. Funkcyjne – wartość jednej zmiennej może być funkcją wartości innej lub innych zmiennych, wartości jednych zmiennych mogą być kombinacjami innych. Oznacza to, że znajomość wartości jednej zmiennej może determinować wartości wielu innych zmiennych. Analiza ta jest bardzo użyteczna nie tylko w kontroli danych, lecz także w imputacji braków danych. 3. Częściowe – zależności, które dotyczą niemal wszystkich jednostek poza wyjątkami. Bada się tu relacje typu: jeżeli → to, z wyjątkiem przypadków {X}. 4. Warunkowe – zależności między wartościami jednej lub wielu zmiennymi zachodzą, o ile spełniony jest określony warunek. Występują tu relacje typu: jeżeli → to: pod warunkiem X.

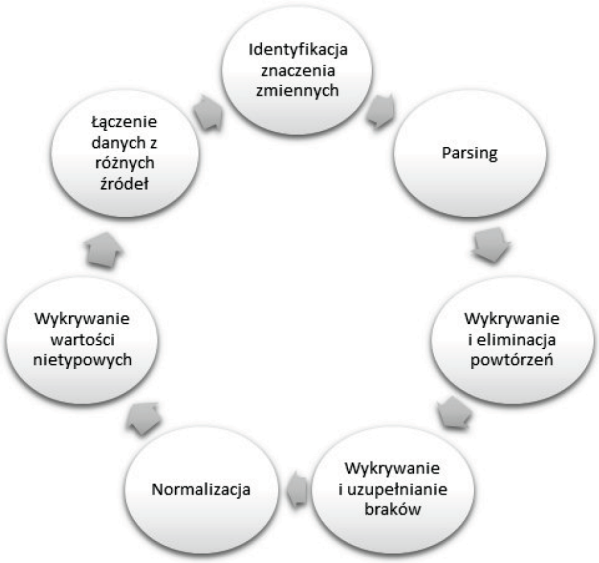
Proces profilowania danych jest niezbędnym, pierwszym etapem przetwarzania *dirty data* w *clear data*. Jednocześnie nie jest procesem pełnym i wystarczającym. Proces profilowania danych nie może zidentyfikować niedokładnych danych, może jedynie wykryć naruszenia zasad czy znanych wzorców. Pozwala wykrywać anomalie, które badacz potrafi sobie wyobrazić. Wiedza uzyskana dzięki profilowaniu danych może być wykorzystana do określenia, jak trudno byłoby wykorzystać istniejące dane do innych celów. Profilowanie może być także wykorzystane do dostarczania danych (lub metadanych) do oceny jakości danych [Batini, Scannapieca 2006]. Może pomóc w określeniu, czy metadane dokładnie opisują dane źródłowe. Proces profilowania danych może być bardzo trudny i kosztowny do praktycznego przeprowadzenia. Wymaga ogromnej wiedzy analityków, a także dostępu do potężnych komputerów. Sama analiza powiązań, dla prostego zbioru danych, złożonego zaledwie ze 100 zmiennych, wymaga dokonania 2^{100}-1 porównań (czyli 1 267 650 600 228 229 401 496 703 205 376 kombinacji kolumn w macierzy danych).

4. Oczyszczenie danych

Drugim etapem procesu przetwarzania *dirty data* w *clear data* jest **oczyszczanie danych** (*data cleansing*). Jest to proces wprowadzania zmian lub usuwania danych w zbiorze danych, które są nieprawidłowe, niekompletne, nieprawidłowo sformatowane lub duplikowane. W jego skład wchodzi szereg podprocesów (rys. 3), które w praktyce powinny być realizowane w sposób ciągły.

Pierwszym etapem jest **identyfikacja znaczenia zmiennych**. Problem ten nie istnieje, jeżeli sami zakładaliśmy zbiór danych. Mamy zdefiniowane zmienne, ustalone skale pomiarowe, mamy świadomość co, dlaczego i w jaki sposób mierzymy. W analizach Big Data często zdarza się, że dane są zbierane przez dużą liczbę różnych podmiotów. Każdy z nich zakłada strukturę danych, która wydaje mu się odpowiednia. Ostatecznie, łącząc dane pochodzące z wielu źródeł, organizowane przez różne instytucje dla różnych celów, otrzymujemy zbiór, którego konstrukcję jest trudno zrozumieć. W szczególności nadal bardzo rzadko zbiór danych jest odpowiednio opisany metadanymi, co w konsekwencji prowadzi do sytuacji, w której analityk nie wie, co faktycznie analizuje. Obserwuje setki kolumn opisanych koda-

mi, których znaczenia może nie znać. W procesie oczyszczania danych etap identyfikacji znaczenia zmiennych trwa bez przerwy.



Rys. 3. Elementy procesu oczyszczania danych

Źródło: opracowanie własne.

Kolejnym etapem oczyszczania danych jest ***parsing***. Jest to proces, w którym dokonywana jest analiza poprawności zapisu danych z punktu widzenia poprawności składni, ortografii i zgodności z przyjętymi standardami. Wykorzystuje się tu wszelkiego rodzaju słowniki, jak i systemy analizy składni. Zasadniczym problemem w procesie parsingu jest dostępność do odpowiednich słowników i reguł gramatycznych. Poza podstawowymi językami europejskimi dostęp do takich materiałów jest ograniczony lub niemożliwy. Jest to podstawowy problem w analizie danych pochodzących z Internetu, który jest siecią globalną, a treści tworzą użytkownicy z dziesiątek krajów posługujący się setkami języków.

W dalszej kolejności na podstawie ustaleń analizy unikatowości dokonuje się **eliminacji duplikatów** ze zbioru analizowanych przypadków. Sam proces eliminacji wydaje się dość prosty, o ile wiadomo, w jaki sposób można duplikaty wykryć – to jednak rola analizy unikatowości.

Podobnie następuje proces **imputacji** braków danych. W analizie kompletności zostają ustalone zasady wykrywania braków danych, a także wykrywania informacji błędnie nadmiarowych. Następnie jedną z wielu znanych metod [Migdał-Najman, Najman 2013] przeprowadza się imputację. W podejściu Big Data wykorzystuje się jednak metody prostsze ze względu na wolumen danych, a także duże wymagania w zakresie szybkości analizy.

Kolejnym etapem oczyszczania danych jest **normalizacja**. Nie chodzi tu o statystyczne metody normalizacji, takie jak np. standaryzacja. Celem tego etapu jest przekształcenie istniejącego zapisu danych do pożądanego formatu, zgodnego z przyjętą normą. O ile w procesie parsingu sprawdzono poprawność ortograficzną i składniową zapisów w zbiorze danych, o tyle w procesie normalizacji jest on przekształcany z postaci surowej (*raw data*) do nowej, zgodnej z przyjętym wzorcem. Przykładem realizowanych tu zadań jest kontrola zapisu osób rejestrujących się na konferencję naukową. Identyfikacja osoby odbywa się na podstawie trzech elementów: 1) imię i nazwisko, 2) tytuł/stopień naukowy, 3) afiliacja. Przykładowa osoba: prof. dr hab. Jan Kowalski, Uniwersytet X może być zapisany w różnych bazach tworzonych na potrzeby dziesiątek konferencji na bardzo wiele sposobów: 1. Jan Kowalski, Kowalski Jan, J. Kowalski, Kowalski J.; 2. Prof. dr hab., Prof. zw. dr hab., Prof. 3. Uniwersytet X, UX, U-X.

Każda kategoria łączy się z każdą, co daje w sumie 4 × 3 × 3 kombinacji tylko w tym prostym przykładzie – a jest to przecież cały czas ta sama osoba. Proces normalizacji porządkuje sposób zapisu danych w poszczególnych rekordach. Zasadniczym problemem jest tu kompletność danych, a także pewność identyfikacji danej jednostki. Ma to szczególne znaczenie przy łączeniu danych pochodzących z różnych źródeł. Brak pełnego imienia lub afiliacji uniemożliwia dokładną identyfikację osoby, a proces integracji zbiorów może stać się problematyczny. Zbudowanie systemu automatycznych powiadomień uczestników konferencji będzie znacznie utrudniony.

Kolejnym etapem jest wykrywanie i ewentualna eliminacja (lub oznaczenie) **wartości nietypowych**. Wiedza o tym, co jest typowe, a co nie w analizowanym zbiorze pochodzi z analizy rozkładu i analizy zakresu. Jeżeli wartości nietypowe zostaną wykryte, należy podjąć decyzję o tym, jakie działanie należy przedsięwziąć: wyeliminować je lub oprzeć dalszą analizę na statystykach odpornych. Warto tu zwrócić uwagę, że w analizie Big Data pojęcie wartości nietypowej jest szczególne. Jeżeli tylko 1:100 000 jednostek jest nietypowa, to w zbiorze setek milionów przypadków są ich tysiące. Może się okazać, że te nietypowe jednostki tworzą wewnętrze spójny segment rynku i jest pożądaną niszą, dla której jesteśmy w stanie stworzyć produkt lub usługę.

Końcowym etapem oczyszczania danych jest ich **integracja**. W wielu przypadkach analizy Big Data końcowy zbiór danych jest w rzeczywistości „zbiorem zbiorów". Informacje o danym podmiocie czy osobie są rozproszone. Część pochodzi z mediów społecznościowych, część z rejestrów (np. sprzedaży) wewnętrznych przedsiębiorstwa, część (którą można kupić) ze zbiorów innych podmiotów. Dane te mają jednak różne formaty, różną strukturę, w różny sposób identyfikują daną jednostkę. W szczególności możemy wiedzieć, że kilka zbiorów opisuje te same jednostki, a jednocześnie może nie istnieć jeden wspólny identyfikator obiektów, pozwalający na ich łatwą integrację. Te dwa problemy: zróżnicowanie struktur i identyfikacja jednostek są w tym etapie kluczowe.

Oczyszczanie danych jest procesem wymagającym dogłębnej wiedzy o analizowanej populacji, branych pod uwagę zmiennych, wymaga dużego nakładu pracy, zaangażowania wielu specjalistów. Proces ten jest trudny do zautomatyzowania, co w konsekwencji powoduje, że jest on organizacyjnie złożony, czasochłonny i kosztowny. W wielu przypadkach oczyszczanie wszystkich danych może nie być uzasadnione finansowo. Koszty takiej pracy mogą być wyższe niż zysk z decyzji podejmowanych na podstawie danych oczyszczonych.

5. Zapobieganie powstawaniu zanieczyszczeń

Trzecim filarem zapewniającym odpowiednią jakość danym w zbiorze Big Data jest zapobieganie powstawaniu błędów (*defect prevention*). Na podstawie poprzednich etapów identyfikuje się przyczyny, źródła, warunki i miejsca powstawania błędów. Planuje i wdraża się następnie mechanizmy zapobiegające ich powstawaniu. Cechą szczególną analizy Big Data jest jednak to, że dane powstają w sposób rozproszony, w wielu miejscach jednocześnie, w wielu często niezgodnych standardach, bardzo często bez jakiejkolwiek kontroli formalnej (Facebook). Żaden administrator czy moderator nie może narzucić tu żadnego standardu, norm czy ograniczeń.

Ograniczając powyższe rozważania do systemów korporacyjnych, można wskazać kilka głównych źródeł zanieczyszczania danych:

1. Błędy w procesie definiowania procesu zbierania, gromadzenia, przechowywania danych – w otwartym środowisku, jakim jest Internet, jedna nieprzemyślana decyzja na etapie projektowania zbioru danych może być trudna do usunięcia, gdy zbiór rośnie o miliony rekordów dziennie.

2. Brak reakcji na wykryte wcześniej problemy – identyfikacja problemu nie jest czynnikiem wystarczającym do podjęcia kroków naprawczych, które mogą być trudne technicznie i bardzo kosztowne. W dalszym jednak horyzoncie jeden problem generuje kolejne, a ich usuwanie jest coraz bardziej kosztowne.

3. Nierozumienie znaczenia różnych elementów danych – w otwartym środowisku mediów społecznościowych, gdy rejestrowana jest każda aktywność internauty, liczba zmiennych opisująca jednego użytkownika może sięgać tysięcy. Trudno opanować logikę wszystkich tych zmiennych i ich wzajemnych powiązań, co prowadzi do automatyzacji działań bez ich zrozumienia.

4. Brak wspólnych metadanych. Problem metadanych jest jednym z istotniejszych. „Dane o danych" są kluczowe do zrozumienia informacji w nich zawartych. Kto?, kiedy?, w jaki sposób?, jak długo?, do kogo?, z kim?, w jakim standardzie?, gdzie?, jakim urządzeniem wprowadził określone dane?, …, pozwala to zrozumieć, co się właściwie stało. Dane te są podstawą procesu integracji danych. Zbieranie metadanych to jednak dodatkowe zadanie, wymagające dodatkowych nakładów finansowych i czasowych. Zwiększa także dodatkowo wolumen zbioru danych.

5. Brak definicji domen – zakresów, dopuszczalnych przedziałów zmienności powoduje kosztowne działania na etapie profilowania i oczyszczania danych.

6. Brak lub istotne błędy w procesie weryfikacji danych. Zaniechania te stoją w sprzeczności z jednym z filarów Big Data, którym jest weryfikacja danych (*veracity*). Jak wskazują badania Craiga Stedmana [2017], jakość danych nie jest nadal w centrum uwagi bardzo wielu firm.

7. Słabe szkolenie pracowników w zakresie wprowadzania danych jest kolejnym źródłem problemów. Wydaje się, że źródło to może być łatwo opanowane, jednak szkolenia to także znaczne koszty. Efekt w postaci poprawy jakości danych nie jest spektakularny i trudno się nim pochwalić np. przed inwestorami.

8. Zaskakującą przyczyną powstawania *dirty data* jest brak motywacji do wprowadzania danych o pożądanej jakości. Brakuje mechanizmów pozwalających bezpośrednio przedstawić zależność między zwiększaniem jakości danych a poprawą efektywności podejmowanych decyzji. Decydentom wydaje się, że wystarczy po prostu dalej zbierać coraz więcej danych. W efekcie końcowym słabo szkoleni pracownicy, bez motywacji do podnoszenia jakości gromadzonych danych, powiększają jedynie wolumen danych. Najczęściej nie są świadomi, że generują *dirty data* i dają pracę analitykom specjalizującym się w oczyszczaniu danych. Koszty są przerzucone do innych działów, których pracę łatwiej uzasadnić.

6. Zakończenie

Analiza Big Data ma wymagania tego samego rodzaju, co każda inna analiza danych. Nadal obowiązuje zasada GIGO (*garbage in, garbage out*): „śmieci na wejściu, to śmieci na wyjściu". Żadna analiza nie może być wartościowa, jeżeli dane, na których się opiera, są niedokładne, częściowo błędne, niepotwierdzone, zawierają liczne duplikaty i braki danych, są niezintegrowane lub nie spełniają wymagań dotyczących standardu zapisu czy kodowania. Znaczny udział *dirty* i *dark data* w zbiorach Big Data niweluje korzyść, jaką można dzięki masowemu zbieraniu i przetwarzaniu danych uzyskać. Profilowanie, oczyszczanie danych i zapobieganie zanieczyszczaniu danych niwelują część tych problemów. Działania te muszą być jednak odpowiednio zorganizowane, zsynchronizowane i uporządkowane w postaci całego systemu złożonego z wielu szczegółowych procedur. Należy pamiętać, że są to działania czasochłonne, wymagają szczególnych i różnorodnych kompetencji od analityków, wymagają specjalnego oprogramowania i znacznych nakładów finansowych. Przy rosnącym wolumenie danych, szybkości ich napływu i rosnących oczekiwaniach co do szybkości uzyskiwania wyników analiz, procedury te są w wysokim stopniu nieefektywne. Coraz częściej przedsiębiorstwa zarabiające dzięki Big Data przyjmują tu postawę pasywną. Zakładają, że wystarczy zwiększać wolumen danych i odsiewać z nich *clear data*, aby osiągać sukces rynkowy. Jeżeli Internet dociera już do kilku miliardów potencjalnych klientów i nawet jeżeli większość danych to „śmieci", to to, co zostanie (*clear data*), to nadal dziesiątki milionów klientów, którzy są w stanie wygenerować satysfakcjonujący zysk.

Literatura

Abedjan Z., Golab L., Naumann F., 2015, *Profiling relational data: a survey*, VLDB Journal, 24, s. 557-581.

Benford F., 1938, *The law of anomalous numbers*, Proceedings of the American Philosophical Society, vol. 78, no. 4, s. 551-572.

Bjornaas K., *6 Quick Dirty Data Stats*, July 27, 2015, https://www.reachforce.com/blog/6-quick-dirty--data-stats/ (20.11.2017).

Cox M., Ellsworth D., 1997, *Application-controlled demand paging for out-of-core visualization*, VIS'97 Proceedings of the 8th Conference on Visualization '97, s. 235.

Luebbe J., 2015, *How Dirty is Social Data? An Analysis of Social Spam*. *Networked Insights*, http://www.networkedinsights.com/socialspam/ (20.11.2017).

Migdał-Najman K., Najman K., 2013, *Samouczące się sztuczne sieci neuronowe w grupowaniu i klasyfikacji danych: teoria i zastosowania w ekonomii*, Wydawnictwo Uniwersytetu Gdańskiego Gdańsk.

Migdał-Najman K., Najman K., 2017, *Big Data = Clear + Dirty + Dark Data*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 469, Taksonomia 29, Klasyfikacja i analiza danych – teoria i zastosowania, s. 131-139.

ReachForce, 2015, *Big Data Marketing, Content Marketing, Marketing Automation*, 25.02.2015, https://www.reachforce.com/blog/is-dirty-data-costing-you-money/ (20.11.2017).

Stedman C., 2017, *Good data quality for analytics becomes an IT imperative*, https://searchdatamanagement.techtarget.com/ehandbook/Good-data-quality-for-analytics-becomes-an-IT-imperative.

Zomaya A.Y., Sakr S. (red.), 2017, *Handbook of Big Data Technologies*, Springer International Publishing AG, Cham, Switzerland.