PAWEŁ KONRAD DOMAGAŁA[1]

BIG DATA I PROFILOWANIE W DOBIE RODO

STRESZCZENIE

Wielkie zbiory danych (Big Data) są wykorzystywane m.in. w telekomunikacji, marketingu, transporcie, motoryzacji, bankowości, turystce i w wielu innych obszarach. Jednakże Big Data poza szansami rozwoju niesie ze sobą pytania natury prawnej. Pojawia się obawa zagrożenia dla autonomii informacyjnej jednostki czy kwestia konieczności wyrażenia zgody na profilowanie. Celem niniejszego opracowania jest analiza pojęcia Big Data i powiązanego z nim profilowania. Wskazane zostaną możliwości wykorzystania Big Data oraz jego ewentualne ograniczenia związane z ochroną danych osobowych.

SŁOWA KLUCZOWE

Big Data, prywatność, autonomia informacyjna

1. Wprowadzenie

Współcześnie wielkie zbiory danych (Big Data) są wykorzystywane m.in. w telekomunikacji, marketingu, transporcie, motoryzacji, bankowości, turystce i w wielu innych obszarach. Jednakże Big Data poza niewątpliwymi szansami rozwoju niesie ze sobą liczne pytania natury prawnej m.in. dotyczące właściwej realizacji zasad przetwarzania danych w wielkich zbiorach. Pojawia się także obawa zagrożenia dla autonomii informacyjnej jednostki[2] czy kwestia konieczności wyrażenia zgody na profilowanie[3]. Współczesna prywatność ulega niewątpliwej

[1] Doktorant na Wydziale Prawa i Administracji Uniwersytetu Kardynała Stefana Wyszyńskiego w Warszawie.
[2] B. FISCHER, *Prawo użytkowników wyszukiwarek internetowych do poszanowania ich autonomii informacyjnej* [w:] *Internet Ochrona wolności, własności i bezpieczeństwa*, red. G. SZPOR, Warszawa 2011, s. 68.
[3] W. WIEWIÓROWSKI, *Profilowanie tylko po powiadomieniu*, Http://www.computerworld.pl/news/377157/GIODO.profilowanie.klientow.tylko.po.powiadomieniu.html.

erozji głównie na skutek decyzji – nie zawsze w pełni świadomych i suwerennych użytkowników Internetu[4]. Coraz częściej zwraca się uwagę na zagrożenia związane z ogromną skalą zbieranych danych, śledzeniem i profilowaniem, bezpieczeństwem, transparentnością a także ich niepoprawnością, możliwościami dyskryminacji czy dużo szerszymi możliwościami prowadzenia nadzoru przez organy różnych państw[5].

Celem niniejszego opracowania jest analiza pojęcia Big Data i powiązanego z nim profilowania. Ponadto wskazane zostaną możliwości wykorzystania Big Data oraz jego ewentualne ograniczenia związane z ochroną danych osobowych. W związku z powyższym Big Data zostanie ukazane w kontekście zmian w obszarze ochrony danych osobowych, które związane były z początkiem stosowania ogólnego rozporządzenia o ochronie danych (RODO)[6].

2. CZYM JEST BIG DATA?

Do tej pory w żadnym akcie prawnym nie zawarto definicji legalnej terminu Big Data. Można jednak wskazać, że odnosi się on do procesu gromadzenia, przetwarzania, analizy danych i wizualizacji wyników z wykorzystaniem wielkich zbiorów danych. Należy w tym miejscu zaznaczyć, że nigdzie nie określa się precyzyjnie, jakiej wielkości ma być omawiany zbiór danych, aby kwalifikował się on jako wielki zbiór danych. To czy dany zbiór danych zostanie określony mianem wielkiego, zależy w głównej mierze od specyfiki konkretnego procesu przetwarzania danych. Dlatego w przypadku procesów, w których w krótkim czasie zbiera się bardzo duże ilości danych, do zakwalifikowania jako wielkiego wymagać się będzie zbioru o wielkości liczonej w Petabajtach (PB). Natomiast w przypadku procesów, w których tempo zbieranych jest dużo mniejsze, do zakwalifikowania danego zbioru jako wielkiego może wystarczyć zbiór o wielkości liczonej jedynie w Megabajtach (MB). W tym miejscu należy wskazać, że niezwykle istotnym, w kontekście kwalifikacji danego zbioru jako Big Data, będzie odpowiedź na pytanie, czy wielkość danego zbioru uniemożliwia dokonywanie analizy zebranych w nim danych z wykorzystaniem tradycyjnych narzędzi analitycznych. W przypadku

[4] I. LIPOWICZ, *Nowe wyzwania w zakresie ochrony danych osobowych* [w:] *Internet Ochrona wolności, własności i bezpieczeństwa*, red. G. SZPOR, Warszawa 2011, s. 14.
[5] P. DROBEK, *Zasada celowości w dobie wielkich zbiorów danych (big data)*, Dodatek do Monitora Prawniczego, Aktualne problemy ochrony danych osobowych 2014, nr 9/2014, s. 21.
[6] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z 27.4.2016 r. W sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (Dz.Urz.UE 2016 Nr L 119/1).

pozytywnej odpowiedzi na powyższe pytanie będziemy mieli do czynienia ze zbiorem Big Data.

Można wskazać, że obecnie do opisywania zbiorów Big Data stosuje się model 5V:

- duża ilość danych (*volume*),
- duża zmienność danych (*velocity*),
- duża różnorodność danych (*variety*),
- duża wiarygodność danych (*velacity*),
- duża wartość danych (*value*).

Powyższy model wskazuje, że do zakwalifikowania danego zbioru jak Big Data konieczne jest spełnienie określonych warunków takich jak: duża ilość danych zawarta w zbiorze, duża zmienność oraz różnorodność danych zawartych w zbiorze, a także duża wiarygodność oraz wartość danych, które składają się na dany zbiór. Jeśli analizowanemu zbiorowi i zawartym w nim danych możemy przypisać powyższe cechy, to będziemy mieli w takiej sytuacji do czynienia ze zbiorem Big Data.

Niezwykle istotną cechą zbiorów Big Data jest to, że zawarte w nich dane pochodzą z różnych źródeł i w dużym stopniu nie są ustrukturyzowane. W omawianym pojęciu nie chodzi tylko o wielkie ilości danych, lecz przede wszystkim o nowe możliwości ich analizy[7]. Pozyskuje się je np. z wyszukiwarek internetowych z portali społecznościowych, chmur obliczeniowych, systemów informacji marketingowej czy baz danych klientów. Podkreślenia wymaga, że dane rzadziej są zbierane przez administratorów w sposób aktywny, lecz raczej są pozyskiwane w sposób pasywny, gdyż są pozostawiane niejako przy okazji przez użytkowników[8].

Należy zaznaczyć, że pojęcie Big Data znacznie ewoluowało i dzisiaj jest postrzegane inaczej, niż jeszcze kilkanaście lat temu. Można wskazać, że pierwsze prognozy dotyczące „boomu informacyjnego", jaki miał nadejść pojawiały się już w latach 40. XX w. Podkreślano wówczas przede wszystkim, problematyczność interpretowania gigantycznych ilości informacji, które ludzkość będzie produkowała w szalonym tempie. Dał temu wyraz F. Rider z Wesleyan University Librarian, publikując w 1944 roku artykuł *The Scholar and the Future of*

[7] Tamże, s. 21.
[8] Tamże, s. 21.

the Research Library. Rider szacował, że wskutek eksplozji danych biblioteki amerykańskich uniwersytetów będą podwajały swoje zbiory średnio co szesnaście lat[9].

Uznaje się, że pierwszy raz terminem Big Data posłużono się 19 lat temu. W sierpniu 1999 r. Bryson, D. Kenwright, M. Cox, D. Ellsworth oraz R. Haimes, wspólnie opublikowali artykuł *Visually exploring gigabyte data sets in real time*[10] i to właśnie w nim, w tytule jednego z podrozdziałów padło określenie Big Data. Wskazano tam, że poprzez Big Data należy rozumieć wielkie zbiory danych, których przetwarzanie, czy dokonywanie ich analiz z wykorzystaniem tradycyjnych narzędzi analitycznych było niemożliwe lub co najmniej wysoce utrudnione.

Natomiast w 2001 roku przedsiębiorstwo analityczno-doradcze specjalizujące się w zagadnieniach strategicznego wykorzystania technologii oraz zarządzania technologiami – Gartner, opublikowało raport[11], podejmujący próbę opisu Big Data poprzez model 3V:

- duża ilość danych (*volume*),
- duża zmienność danych (*velocity*),
- duża różnorodność danych (*variety*).

Ponadto w słowniku udostępnionym przez Gartner wskazano, że *Big Data to duże zasoby informacyjne o dużej przepustowości i dużej różnorodności wymagające opłacalnych, innowacyjnych form przetwarzania w celu lepszego podejmowania decyzji*[12].

W tym miejscu należy zaznaczyć, że pojęcie Big Data tak samo jak sposoby przetwarzania danych ulegało znacznym zmianom. Głównym ich powodem był znaczący przyrost ilości zbieranych danych, a także ich wysoka zmienność w krótkim czasie. Powyższe powodowało, że duża cześć zbieranych danych w krótkim czasie stawała się nieaktualna i prowadziła do zakłamywania wyników sporządzanych analiz.

W związku z powyższym w pewnym momencie Big Data, ze względu na nieaktualność przetwarzanych danych zaczęto określać mianem Bug Data, czyli danymi śmieciowymi. W tym

[9] *Historia w pigułce: big data*, http://www.brief.pl/artykul,2824,historia_w_pigulce_big_data.html.
[10] S. BRYSON, D. KENWRIGHT, M. COX, D. ELLSWORTH, R. HAIMES, *Visually exploring gigabyte data sets in real time*, Communications of the ACM, nr 42/8, s. 82-90.
[11] D. LANCY, *3D Data Management Controlling Data Volume Velocity and Variety*, Stamford 2001, https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
[12] *Big data*, https://www.gartner.com/it-glossary/big-data.

samym czasie opisane zjawisko przetwarzania nieaktualnych danych zaczęto określać mianem ROT, czyli *Redundant, Obsolete and Trivial*, co w tłumaczeniu oznacza dane zbędne, przestarzałe oraz nieważne. Dlatego wspomniany model 3V rozszerzono o dwie dodatkowe składowe – dużą wiarygodność (*velacity*) i dużą wartość (*value*) gromadzonych danych, tworząc w ten sposób obecnie wykorzystywany model 5V[13].

Przedstawiona zmiana w znacznej mierze zwiększyła użyteczność wykorzystania Big Data. Ponadto charakter i ewolucja składowych modeli opisujących Big Data dobitnie pokazuje wzrost znaczenia na przestrzeni dekady analizy wielkich zbiorów danych. Ponadto ostatnimi czasy w obszarach, które cechowały się szczególnie dużą zmiennością, a więc krótkim czasem po jakim dane stawały się nieaktualne, w celu ograniczania dokonywania analiz, w oparciu o nieaktualne dane zebrane w ramach Big Data, zaczęto wykorzystywać jedynie dane, które zostały pozyskane w czasie ostatniego roku, czy nawet półrocza. Natomiast takie zbiory Big Data zaczęto określać mianem zbiorów Fast Data.

3. BIG DATA JAKO PROCES

W celu zrozumienia sposobu przetwarzania danych należałoby wyjaśnić pewne pojęcia, które są wykorzystywane w procesie przetwarzania danych w zbiorach Big Data i są z nimi nierozerwanie związane. Pierwszym pojęciem ściśle związanym ze zbiorami Big Data jest *Data science*. Termin ten można rozumieć jako naukę o danych, traktującą o procesie pozyskania, obróbki, wizualizacji i wnioskowania w oparciu o dane ustrukturyzowane i nieustrukturyzowane, z użyciem metod statystycznych, eksploracji danych, uczenia maszynowego i analizy predykcyjnej. Ściśle powiązane z powyższym terminem jest pojęcie *Data scientist*, czyli osoby zajmującej się analizą danych nieuporządkowanych, które wchodzą w skład wielkich zbiorów danych typu Big Data.

Kolejnym niezwykle istotnym pojęciem powiązanym z Big Data jest *Data mining*. Pojęcie to często określa się mianem eksploracji lub wydobywania danych. Można wskazać, że *Data mining* jako obszar zainteresowania informatyki jest znany już od bardzo wielu lat. Omawiana technika polega na pozyskiwaniu nowych informacji, czy wiedzy z posiadanych już wcześniej zbiorów danych. Dlatego można wskazać, że *Data mining* to przede wszystkim zbiór

[13] A. MEDNIS, *Big data a regulacje prawne*, Warszawa 2014, s. 5.

nowoczesnych technik analitycznych, które pozwalają na zautomatyzowane odkrywanie statystycznych zależności i schematów w dużych zbiorach danych. Poznanie wcześniej nieznanych zależności i schematów, przedstawianych następnie w formie reguł logicznych, drzew decyzyjnych czy sieci neuronowych może posiadać dużą wartość ekonomiczną i może zostać użyte do wspomagania podejmowania decyzji finansowych i marketingowych w przedsiębiorstwie. Często środowiska oparte o *Data mining* wykorzystują zaawansowane algorytmy uczenia maszynowego oraz duże zbiory danych. Jednak mimo że eksploracja danych wyrosła na gruncie sztucznej inteligencji i wspomnianego uczenia maszynowego, to jednak rozmiary stawianych przed nią problemów skutkuje koniecznością opracowania zupełnie nowych, wyrafinowanych algorytmów, metod i architektur. Można ponadto wskazać, że celem zastosowania omawianych technik *Data mining* jest lepsze wykorzystanie danych, które dany podmiot już posiada i pozyskanie z nich informacji, które mogą przyczynić się do zmniejszenia kosztów działania lub zwiększenia zysków z oferowanych na rynku produktów lub usług.

Wspomniane już wcześniej uczenie maszynowe (*Machine learning*) jest nierozerwalnie związane z pojęciem *Data mining*, gdyż właściwie każdy model wykorzystywany w *Data mining* opiera się na uczeniu maszynowym tj. na uczeniu się przez program lub algorytm wykrywać pewne zależności pomiędzy zgromadzonymi danymi. Uczenie maszynowe może być realizowane w oparciu o dwa różne modele. Pierwszy z nich to *supervised machine learning*, czyli sytuacja, kiedy program otrzyma na wstępie określony materiał do nauczenia. Natomiast drugi model to *unsupervised machine learning*, gdzie zakłada on, że program nie otrzyma na początku materiału do nauczenia.

Cykl życia informacji przetwarzanej w skali Big Data polega na przetwarzaniu jej w różnych typach następujących po sobie procesów. W zależności od przyjętej metodologii, odmienne mogą być rodzaje oraz ilość etapów składających się na cały proces[14]. Niezależnie od przyjętego modelu cel pozostaje tożsamy dla wszystkich z nich – osiąganie lepszych rezultatów na różnych płaszczyznach. Analiza Big Data może służyć zwiększeniu wydajności, minimalizacji błędów, zmniejszeniu kosztów czy optymalizacji procesów.

Na potrzeby niniejszego opracowania przyjęty został cykl życia Big Data, obejmujący cztery następujące po sobie fazy tj. generowanie danych, akwizycję danych, przechowywanie danych i analizę danych.

[14] Por. *Big data lifecycle*, https://cloud.google.com/solutions/data-lifecycle-cloud-platform.

3.1. GENEROWANIE DANYCH

Generowanie danych dotyczy tego, w jaki sposób wytwarzane są dane przetwarzane na dużą skalę, które zawierają różnorodne i złożone zestawy informacji z różnych heterogenicznych lub rozproszonych źródeł. Źródła pozyskiwania danych to np. rejestry publiczne, portale społecznościowe, nadajniki GPS, video-monitoring, telefony komórkowe, smartwatch'e itp. Jako typowe źródła pochodzenia danych wskazuje się sektor biznesowy, Internet oraz badania naukowe[15].

Niezależnie od źródła pochodzenia, dane można podzielić na trzy rodzaje w zależności od stopnia ich uporządkowania[16]:

- dane ustrukturyzowane,
- dane nieustrukturyzowane,
- dane częściowo ustrukturyzowane.

Często trudno jest ustalić, w jakim stopniu dane są uporządkowane. Dane w rejestrach publicznych zwykle np. KRS czy CEIDG uznaje się za ustrukturyzowane, ale już dane znajdujące się na skrzynce e-mailowej będą ustrukturyzowane tylko częściowo (dodatkowo stopień ich uporządkowania w ramach skrzynek e-mailowych oferowanych przez kilku dostawców poczty może się od siebie różnić). Zwykle pomocne tu będzie subiektywne kryterium, związane ze stopniem problematyczności odszukania danej informacji. Im wyższy stopień trudności w odszukaniem danego rekordu, lub z porównaniem różnych rekordów w ramach bazy danych biorąc pod uwagę różne kryteria, tym bardziej prawdopodobne jest, że administrator przetwarza dane w nieustrukturyzowanej bazie.

3.2. AKWIZYCJA DANYCH

Akwizycja, czyli zbieranie danych pochodzących z różnych źródeł, polega na ich przetwarzaniu w celu dalszej analizy (w tym ustalenia zależności pomiędzy nimi czy uzyskania zupełnie nowych informacji). Czasami akwizycji może towarzyszyć także filtrowanie danych

[15] K. VENKATARAMANAN, M. SREEDEVI, *A Review on Big Data Concepts*, International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 3 (2016), s. 278.
[16] E. THOMAS, *Big Data Fundamentals Concepts, Drivers & Techniques*, Crawfordsville 2015.

pozwalające na ustalenie, które dane są istotne dla dalszej analizy, a także wykluczenie danych nieistotnych, nieaktualny lub niepełnych[17].

3.3. PRZECHOWYWANIE DANYCH

W ramach cyklu przetwarzania danych w skali Big Data, na etap przechowywania składa się trwałe utrzymywanie danych (np. na serwerach, w magazynach) i zarządzanie nimi. W systemie przechowywania danych jest więc istotna zarówno infrastruktura sprzętowa, jak i zarządzanie danymi. Infrastrukturę sprzętową tworzą wspólne zasoby teleinformatyczne zorganizowane w taki sposób, aby pozwalały na wykonywanie wielu zróżnicowanych zadań.

Etap przechowywania jest czasami nazywany etapem „przygotowania" ponieważ skupia się m.in. na normalizacji zestawów danych, uzgodnieniu formatów dat i systemów współrzędnych geograficznych, usunięcia duplikatów, podziału kolumn, nagłówków dostaw i ogólnie uczynienia zestawu danych użytecznym dla programu analitycznego[18].

3.4. ANALIZA DANYCH

Zasadniczym etapem przetwarzania danych w ramach Big Data jest ich analiza. Analizę można podzielić na 6 kluczowych obszarów technicznych: analizę danych strukturalnych, analizę tekstu, analizę multimedialną, analitykę internetową, analizę sieci oraz analizę mobilną. Klasyfikacja ta ma na celu podkreślenie kluczowych cech danych analizowanych w ramach każdego obszaru[19].

Sama analityka danych to dziedzina, zajmująca się sprawowaniem kontroli pełnym cyklem życia danych, który obejmuje zbieranie, czyszczenie, organizowanie, przechowywanie, analizowanie i zarządzanie danymi. Z terminie tym są ponadto związane zagadnienia rozwoju metod analizy, technik badawczych i automatycznych narzędzi. W środowiskach wielkich zbiorów danych wyzwaniem jest opracowanie i korzystanie z wysoce skalowalnych

[17] A. LABRINIDIS, H. JAGADISH, *Challenges and opportunities with Big Data*, Proceedings of the VLDB Endowment, 5/2012, s. 2032.
[18] L. POUCHARD, *Revisiting the Data Lifecycle with Big Data Curation*, International Journal of Digital Curation 2015, Vol. 10 Issue 2 (2015), s. 186.
[19] K. VENKATARAMANAN, M. SREEDEVI, op. Cit., s. 278.

rozproszonych technologii i struktur, które są w stanie analizować duże ilości danych pochodzące z różnych źródeł. Analiza dużych zbiorów danych różni się od tradycyjnej analizy danych przede wszystkim ze względu na objętość, prędkość i różnorodność przetwarzanych danych. Podczas analizy danych dochodzi do wielu etapów przetwarzania. W pierwszej kolejności następuje ustalenie kontekstu biznesowego, w ramach którego ma nastąpić analiza. Kolejnym etapem jest identyfikacja danych, które okażą się użyteczne z perspektywy analizowanego przypadku biznesowego. W dalszej kolejności następuje pozyskanie danych oraz ich filtrowanie pod kątem przydatności, aktualności czy prawidłowości[20].

Czwarty etap polega na wydobyciu i wyodrębnieniu danych. Niektóre dane zidentyfikowane jako dane wejściowe do analizy mogą przybrać format niezgodny z rozwiązaniem stosowanym w procesie analizy Big Data. Nieprawidłowe dane mogą zniekształcać i w konsekwencji sfałszować wyniki analizy. W przeciwieństwie do tradycyjnych danych pozyskiwanych przez przedsiębiorstwa (np. dane od klienta wypełniającego formularz zamówienia), gdzie struktura danych jest domyślnie zdefiniowana, a dane są wstępnie zatwierdzane, dane wprowadzane do analiz Big Data mogą być nieustrukturyzowane bez oraz niesprawdzone pod kątem ich prawidłowości. Dlatego ten etap jest przeznaczony na ustalanie złożonych reguł pozwalających na sprawdzanie poprawności i usuwanie wszelkich zidentyfikowanych nieprawidłowości. Ponadto wyodrębnianie danych służy nie tylko ich identyfikacji, ale także przekształceniu ich w format, który stosowane oprogramowanie analityczne może wykorzystać do celów analizy danych. Zakres wymaganej ekstrakcji i transformacji jest każdorazowo inny i zależy od rodzaju analityki i możliwości rozwiązania Big Data.

W dalszym stadium następuje agregacja danych i wyłonienie rekordów reprezentatywnych. Z uwagi na to, że dane mogą być rozlokowane w wielu bazach wymaga to łączenia zestawów danych za pośrednictwem wspólnych pól, na przykład daty lub identyfikatora. Konieczne jest uzgodnienie danych lub określenie zestawu danych reprezentujących prawidłową wartość. Etap agregacji danych i reprezentacji jest poświęcony integracji wielu zestawów danych i scalenia wyników tej integracji. Wykonanie tego etapu może się skomplikować z powodu różnic w:

[20] Ibidem, s. 279.

- strukturze danych – mimo, że format danych może być taki sam, model danych może być inny.
- semantyce - wartość oznaczona inaczej w dwóch różnych zestawach danych może oznaczać to samo, na przykład „miejsce zamieszkania" i „adres zamieszkania".

Skala na jaką przetwarzane są dane może powodować, że agregacja danych będzie wymagać dużo czasu i wysiłku. Uzgodnienie tych różnic może wymagać złożonej metodyki, która będzie wykonywana automatycznie, bez potrzeby interwencji człowieka. To właśnie na tym etapie należy dokonać założeń dla wymagań dotyczących analizy danych, aby pomóc w zwiększeniu ponownego wykorzystania danych[21].

Po etapie agregacji następuje etap właściwej analizy. Ten etap może mieć charakter iteracyjny, zwłaszcza jeśli analiza danych ma charakter eksploracyjny, w którym to przypadku powtarza się analizę aż do wykrycia odpowiedniego wzoru lub korelacji. W zależności od wymaganego wyniku analitycznego ten etap może być tak prosty, jak wysłanie zapytania do zestawu danych w celu obliczenia agregacji dla porównania. Z drugiej strony może to być bardzo złożone działanie, jak połączenie eksploracji danych i złożonych technik analizy statystycznej w celu wykrycia wzorców i anomalii lub wygenerowania modelu statystycznego lub matematycznego do przedstawienia zależności między zmiennymi.

Przedostatnim etapem procesu analizy Big Data jest wizualizacja wyników. Możliwość analizowania ogromnych ilości danych i znajdowania użytecznych statystyk niesie małą wartość, jeśli jedynymi, którzy potrafią zinterpretować wyniki, są analitycy. Ten etap jest poświęcony wykorzystaniu technik wizualizacji danych i narzędzi do graficznego przekazywania wyników analiz, które pozwalają na efektywną ich interpretacji przez sektor biznesowy. Wyniki ukończenia etapu wizualizacji danych zapewniają użytkownikom możliwość przeprowadzenia analizy wizualnej, pozwalając na odkrycie odpowiedzi na pytania, których jeszcze nie sformułowali użytkownicy. Należy zauważyć, że same wyniki mogą być prezentowane na wiele różnych sposobów, co może wpływać na sposób ich interpretacji. Prezentacja wyników dla samej prezentacji nie miałaby jednak sensu. Cały etap analizy Big Data prowadzi w końcu do wykorzystania wyników w praktyce. Bazując na wynikach analizy danych, użytkownicy systemu gwarantowania i rozstrzygania roszczeń rozwinęli wiedzę o charakterze fałszywych roszczeń. Aby jednak uzyskać wymierne korzyści z tej analizy

[21] Ibidem, s. 280.

danych, generowany jest model oparty na technice uczenia maszynowego, który jest następnie włączany do istniejącego systemu przetwarzania roszczeń w celu oznaczania fałszywych twierdzeń[22].

4. BIG DATA A PROFILOWANIE

Pojęcie Big Data jest nierozerwalnie związane z zagadnieniem profilowania, które jest zjawiskiem dość nowym, a w doktrynie i orzecznictwie istnieje stosunkowo niewielka ilość wyczerpujących opracowań z nim związanych. Do tej pory budzi wiele obaw i zastrzeżeń co widoczne jest nawet na szczeblu unijnym – wątpliwości i zalecenia związane z profilowaniem wyraził w 2010 roku w swojej rekomendacji Komitet Ministrów Państw Członkowskich[23].

Na bardzo ogólnym poziomie, profilowanie można porównać do kategoryzowania osób na podstawie różnych cech. Zarówno tych „niezmiennych" (np. płeć, pochodzenie etniczne, wiek, kolor oczu) jak i „zmiennych" (zachowanie, zwyczaje, preferencje)[24]. Zazwyczaj profile tworzy się za pomocą techniki zwanej „analizą behawioralną". Polega ona na dopasowaniu i korelacji określonego zachowania (np. wyborów konsumenckich) z cechami (np. wiek).

W poprzednio obowiązującej Ustawie o ochronie danych osobowych z 1997 r. brakowało przepisu, który stanowiłby legalną definicję tego pojęcia. Jednakże zawierała ona pewne regulacje odnoszące się do profilowania. W art. 26a przytoczonej ustawy sformułowano generalny zakaz wydawania ostatecznych rozstrzygnięć w indywidualnej sprawie, jeśli treść tego rozstrzygnięcia jest wyłącznie wynikiem operacji na danych osobowych, prowadzonych w systemie informatycznym. Definicja profilowania została zawarta natomiast w przepisach obecnie obowiązującego RODO. W jego art. 4 pkt 4 wskazano że: *profilowanie oznacza dowolną formę zautomatyzowanego przetwarzania danych osobowych, które polega na wykorzystaniu danych osobowych do oceny niektórych czynników osobowych osoby fizycznej, w szczególności do analizy lub prognozy aspektów dotyczących efektów pracy tej osoby*

[22] Ibidem, s. 284.
[23] *Rekomendacja CM/Rec (2010) 13 Komitetu Ministrów państw członkowskich w sprawie ochrony osób w związku z automatycznym przetwarzaniem danych osobowych podczas tworzenia profili*, Strasburg 2010.
[24] *Podniesienie skuteczności działa policji. Rozumienie dyskryminującego profilowania etnicznego i zapobieganie mu: przewodnik*, Luksemburg 2010, s. 8.

fizycznej, jej sytuacji ekonomicznej, zdrowia, osobistych preferencji, zainteresowań, wiarygodności, zachowania, lokalizacji lub przemieszczania się.

Wskazuje się (mimo że nie jest to jasno w przepisie wyrażone), że przetwarzanie musi określać osobowość pewnej jednostki na bazie pewnego profilu standardowego i tą drogą udostępnić zautomatyzowane rozstrzygnięcie[25]. Jako typowy przykład takiego rodzaju decyzji w nauce prawa podaje się sytuacje, w których w bazie danych zawierającej dane osobowe zapisane zostały pewne kategorie informacji. Następnie specjalistyczne oprogramowanie, obejmujące pewien algorytm przetwarzania informacji wykonuje operacje na danych osobowych, wynikiem których jest w każdym wypadku podjęcie decyzji, odnoszącej się od osób, których dane są przetwarzane. W praktyce tego rodzaju sytuacje występować mogą stosunkowo często w związku z analizą zdolności kredytowej osób fizycznych w oparciu o podane informacje[26].

Na podstawie przepisów o ochronie danych osobowych obowiązujących przed 25 maja 2018 r. wskazywano, że wydawanie rozstrzygnięć zawierających ocenę osoby i przez to dotykających jej praw osobistych nie powinno być przekazywane komputerom; za tego rodzaju decyzje powinien być zawsze odpowiedzialny człowiek. Przyjęte w dyrektywie i polskiej ustawie o ochronie danych osobowych rozwiązania odpowiadały koncepcji *prawa do informatycznego samookreślenia się* i sprzeciwia się ignorowaniu indywidualności jednostki ludzkiej, degradowaniu jej do roli obiektu komputerowych operacji[27].

Profilowanie jako jedna z form przetwarzania danych została szerzej opisana w RODO, które 25 maja 2018 r. zastąpiło Dyrektywę 95/46/WE[28]. Zgodnie z treścią rozporządzania, za profilowanie uważa się *dowolną formę zautomatyzowanego przetwarzania danych osobowych, które polega na wykorzystaniu danych osobowych do oceny niektórych czynników osobowych osoby fizycznej, w szczególności do analizy lub prognozy aspektów dotyczących efektów pracy tej osoby fizycznej, jej sytuacji ekonomicznej, zdrowia, osobistych preferencji, zainteresowań, wiarygodności, zachowania, lokalizacji lub przemieszczania się*. W uzasadnieniu do RODO wskazuje się, że *należy poinformować osobę, której dane dotyczą, o fakcie profilowania oraz o konsekwencjach takiego profilowania. Jeżeli gromadzi się dane osobowe od osoby, której*

[25] J. BARTA, P. FAJGIELSKI, R. MARKIEWICZ, *Ochrona danych osobowych Komentarz*, Warszawa 2015, s. 483.
[26] D. BAINBRIDGE, *EC Data Protection Directive*, Londyn – Dublin – Edynburg 1996, s. 67-68.
[27] J. BARTA, P. FAJGIELSKI, R. MARKIEWICZ, op. Cit., s. 480-481.
[28] Dyrektywa 95/46/WE Parlamentu Europejskiego i Rady z 24.10.1995 r. W sprawie ochrony osób fizycznych w zakresie przetwarzania danych osobowych i swobodnego przepływu tych danych (Dz.Urz.WE 1995 Nr L 281/31).

dane dotyczą, należy ją też poinformować, czy ma ona obowiązek je podać, oraz o konsekwencjach ich niepodania. Ponadto w uzasadnieniu wskazano, że *podejmowanie decyzji na podstawie takiego przetwarzania, w tym profilowania, powinno być dozwolone, w przypadku gdy jest to wyraźnie dopuszczone prawem Unii lub prawem państwa członkowskiego, któremu podlega administrator*. Wielokrotne odwołania do profilowania w treści RODO dają podstawy sądzić, że zagadnienie to wymaga coraz większej uwagi i uregulowania w prawie, gdyż jego znaczenie jest niebagatelne a wykorzystanie profilowania w biznesie (zarówno w sektorze publicznym jak i prywatnymi) wzrasta z każdym rokiem.

5. PODSUMOWANIE

Ze zbiorów Big Data płyną oczywiste korzyści dla wielu sektorów gospodarki. Można je wykorzystywać w transporcie publicznym, medycynie, genetyce, ekologii, transporcie publicznym, nauce, rozwijaniu internetu rzeczy (*Internet of Things* - IoT), optymalizacji sprzedaży, zarządzania personelem itp. Przede wszystkim są wykorzystywane przez firmy w działaniach reklamowych ponieważ pozwalają na bardziej ukierunkowany i skuteczny marketing. W bardzo ciekawy sposób Big Data zostało wykorzystane w przemyśle browarniczym. Izraelski startup WeissBeerger, jako pierwszy na świecie opracował i wdrożył w barach, pubach i restauracjach system monitorujący na żywo: ilość, gatunek i markę spożywanego piwa, mierzone w konkretnych przedziałach czasowych. Zlicza także przychody oraz pozostałe na zapleczu beczki (kegi) piwa. Dzięki temu właściciele pubów wiedzą, jakie marki i o jakich godzinach „piją się najlepiej". Wspomniany system również dostarcza właścicielom instrukcje SMS, podpowiadające jakie piwo, w jakich ilościach i w jakie dni powinni zamawiać[29]. Innym przykładem zastosowania Big Data były działania podjęte przez spółkę Google, która, nie tylko w swojej wyszukiwarce wykorzystuje analizę wielkich zbiorów danych, ale także sponsoruje badania nad rozwiązaniem globalnego problemu malarii w Afryce[30].

Należy wskazać, że mimo iż zjawisko wielkich baz danych budzi sporo obaw, to w świetle obowiązujących przepisów przetwarzanie danych nawet na tak ogromną skalę nie

[29] *8 zastosowań Big Data, o których nie miałeś pojęcia*, https://www.focus.pl/artykul/8-zastosowan-big-data-o-ktorych-nie-miales-pojecia.
[30] *11 najfajniejszych zastosowań Big Data*, https://plblog.kaspersky.com/10-najfajniejszych-zastosowan-big-data/2812/.

jest zakazane. Warunkiem jednak jest, że zostaną spełnione obowiązki wynikające z RODO, zwłaszcza obowiązek informacyjny i nie zostaną naruszone zasady przetwarzania danych. W szczególności podkreśla się problematyczność sprostania wymogowi celowości[31] i adekwatność zbieranych i przetwarzanych danych.

***

BIG DATA AND PROFILING VERSUS GDPR

Nowadays, large data sets (Big Data) are used in telecommunications, marketing, transport, automotive, banking, tourism and many other areas. Nonetheless, aside of obvious opportunities for development, Big Data triggers numerous questions of a legal nature, including the proper implementation of the principles of data processing in large collections. There is also concern about the threat to the information autonomy of the individual or the need to give consent to profiling. Privacy is eroding mainly due to decisions taken by not always fully conscious and sovereign Internet users. The risks are increasingly recognised, stemming from the huge scale of data collection, tracking and profiling, security, transparency, and their incorrectness, discrimination and much wider supervision by the authorities of various countries.

This paper examines the concept of Big Data and related profiling. Moreover, Big Data application and its possible limitations linked to personal data protection will be discussed. Accordingly, Big Data will be presented amidst the changes in personal data protection, once the General Data Protection Regulation (GPRD) has come into force.

BIBLIOGRAFIA

11 najfajniejszych zastosowań Big Data, https://plblog.kaspersky.com/10-najfajniejszych-zastosowan-big-data/2812/

[31] P. DROBEK, op. Cit., s. 21.

8 zastosowań Big Data, o których nie miałeś pojęcia, https://www.focus.pl/artykul/8-zastosowan-big-data-o-ktorych-nie-miales-pojecia

BAINBRIDGE D., *EC Data Protection Directive*, Londyn – Dublin – Edynburg 1996.

BARTA J., FAJGIELSKI P., MARKIEWICZ R., *Ochrona danych osobowych Komentarz*, Warszawa 2015.

BIELAK-JOMAA E., LUBASZ D., RODO, *Ogólne Rozporządzenie o Ochronie Danych, Komentarz*, Warszawa 2017.

Big data lifecycle, https://cloud.google.com/solutions/data-lifecycle-cloud-platform.

Big data, https://www.gartner.com/it-glossary/big-data.

BRYSON S., KENWRIGHT D., COX M., ELLSWORTH D., HAIMES R., *Visually exploring gigabyte data sets in real time*, Communications of the ACM, nr 42/8.

DROBEK P., *Zasada celowości w dobie wielkich zbiorów danych (big data)*, Dodatek do Monitora Prawniczego, Aktualne problemy ochrony danych osobowych 2014, nr 9/2014.

FISCHER B., *Prawo użytkowników wyszukiwarek internetowych do poszanowania ich autonomii informacyjnej* [w:] *Internet Ochrona wolności, własności i bezpieczeństwa*, red. G. SZPOR, Warszawa 2011.

Historia w pigułce: big data, http://www.brief.pl/artykul,2824,historia_w_pigulce_big_data.html.

LABRINIDIS A., JAGADISH H., *Challenges and opportunities with Big Data*, Proceedings of the VLDB Endowment, 5/2012.

LANCY D., *3D Data Management Controlling Data Volume Velocity and Variety*, Stamford 2001, https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

LIPOWICZ I., *Nowe wyzwania w zakresie ochrony danych osobowych* [w:] *Internet Ochrona wolności, własności i bezpieczeństwa*, red. G. SZPOR, Warszawa 2011.

LITWIŃSKI P., BARTA P., KAWECKI M., *Rozporządzenie UE w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i swobodnym przepływem takich danych. Komentarz*, Warszawa 2017.

MEDNIS A., *Big data a regulacje prawne*, Warszawa 2014..

Podniesienie skuteczności działa policji. Rozumienie dyskryminującego profilowania etnicznego i zapobieganie mu: przewodnik, Luksemburg 2010

POUCHARD L., *Revisiting the Data Lifecycle with Big Data Curation*, International Journal of Digital Curation 2015, Vol. 10 Issue 2 (2015).

Rekomendacja CM/Rec (2010) 13 Komitetu Ministrów państw członkowskich w sprawie ochrony osób w związku z automatycznym przetwarzaniem danych osobowych podczas tworzenia profili, Strasburg 2010..

THOMAS E., *Big Data Fundamentals Concepts, Drivers & Techniques*, Crawfordsville 2015

VENKATARAMANAN K., SREEDEVI M., *A Review on Big Data Concepts*, International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 3 (2016).

WIEWIÓROWSKI W., *Profilowanie tylko po powiadomieniu*, http://www.computerworld.pl/news/377157/GIODO.profilowanie.klientow.tylko.po.powiadomieniu.html.

Dyrektywa 95/46/WE Parlamentu Europejskiego i Rady z 24.10.1995 r. w sprawie ochrony osób fizycznych w zakresie przetwarzania danych osobowych i swobodnego przepływu tych danych (Dz.Urz.WE 1995 Nr L 281/31).

Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z 27.4.2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (Dz.Urz.UE 2016 Nr L 119/1).