CZĘŚĆ I
Analityka danych – zagadnienia podstawowe

Rozdział 1.
Cele analityki danych

W świecie *Big Data*, w którym każda organizacja może uzyskać zdalny dostęp do ogromnych zasobów danych, zasada „gdy wrzucisz śmieci do systemu, to końcowym efektem ich przetwarzania będą również śmieci", nabiera nowego znaczenia. Skupiając się na rozwoju narzędzi analitycznych i możliwościach ich użycia zdajemy się zapominać, że podstawą dla prawidłowej oceny świata poddawanego analizie jest poprawne określenie, jakie dane i dla jakiego celu chcemy przetwarzać.

Rozważania dotyczące roli analityki danych należy zacząć od omówienia oczekiwań wobec niej. Celem wykorzystywania narzędzi *Big Data* jest identyfikowanie takich zachowań indywidualnych osób, które przynoszą mniej zagrożeń i potencjalnie kreują większe przychody dokonującemu analizy podmiotowi[1]. Wszelkie odstępstwa od modelu zachowania uważanego za najbardziej akceptowany są traktowane jako z założenia podejrzane. Taki efekt analiz *Big Data* może paradoksalnie być największym niebezpieczeństwem dla samego modelu. Oznacza on bowiem, że wszelka odmienność i innowacyjność jest traktowana jako potencjalne zagrożenie. System informacyjny zaczyna działać, jak – znany ze starej piosenki – „centralny wyrównywacz". Może to prowadzić do szczególnych form dyskryminacji w sektorze bankowym czy ubezpieczeniowym, ale również prowadzić musi do podobnych skutków w sektorze publicznym. Przy działaniach dokonywanych na podstawie danych z zasobów prywatnych i z rejestrów publicznych należy przede wszystkim ocenić jakość danych w nich zawartych i spójność pozyskiwanego z danego zasobu „wyciągu danych"[2].

[1] *V. Mayer-Schoenberger*, *K. Cukier*, Big Data. Rewolucja, która zmieni nasze myślenie, pracę i życie, Warszawa 2014, s. 170–171, w ciekawy sposób opisują, dlaczego niektórych badań nie mogą i nie chcą prowadzić administratorzy pierwotnych danych i dlaczego tym samym można stworzyć model biznesowy dla ponownego wykorzystania tych danych.

[2] O problemie automatyzacji błędnych decyzji lub decyzji opartych na niereprezentatywnych danych pisze *V. Eubanks*, Automating Inequality. How high-tech tools profile, police and punish the poor,

Zagrożenie zmianą podstawowych zachowań społecznych w związku z ryzykiem wykorzystania danych przeciwko nam samym staje się bardzo poważne, jeśli w jakimkolwiek momencie miałoby dojść do ponownego przetwarzania indywidualnych danych dotyczących oświaty czy ochrony zdrowia w modelu otwartych danych. Już sama świadomość, że nasze zachowanie jest stale obserwowane przez „inteligentną szkołę", „inteligentne e-zdrowie" czy „inteligentne miasto", które mogą się później swą wiedzą o nas dzielić z innymi, będzie ziszczeniem się obaw przed „nowym wspaniałym światem", w którym wszyscy się nawzajem obserwują[3].

W czasach *Big Data* na pierwszy plan wysuwa się jednak inne oczekiwanie wobec nowoczesnych narzędzi analitycznych przetwarzających ogromne zasoby danych. Profesor *Bolesław Szafrański*, poszukując polskiego tłumaczenia terminu *Big Data*, proponował używanie sformułowania „moc danych"[4], ponieważ z jednej strony określa ono, że tych danych „jest moc", czyli że mamy do czynienia z mnóstwem ogromnych zasobów, z drugiej strony te zasoby i to co w danych się znajduje, kreują – poprzez efekt synergii – dodatkową moc dla wszystkich, którzy je przetwarzają[5].

Kolejne etapy informatyzacji procesów gospodarczych i administracyjnych oraz rosnąca rola środków komunikacji elektronicznej w życiu codziennym powodowały, że już od lat 70. wzrastało przekonanie, że istnieją podmioty, które uzyskują bądź mogą uzyskać dostęp do lawinowo rozwijających się zasobów informacyjnych i mogą wdrożyć sposoby przetwarzania informacji, które – nieznane nieświadomym uczestnikom rynku i obywatelom – mogą prowadzić do podejmowania wobec nich środków, których nie są świadomi i które wręcz mogą prowadzić do dyskryminacji osób, grup społecznych czy przedsiębiorców. Pierwotnie organizacją, która była w naturalny sposób oskarżana o chęć świadomego, acz ukrytego przetwarzania rozproszonych danych w sposób, który może naruszać prawa i wolności, było państwo. Szybko rozprzestrzeniało się przekonanie, że praktyki potężnych rządów i korporacji w zakresie przetwarzania danych redukują jednostki do statusu przedmiotu danych, co zagraża prawom podstawowym i wolnościom. Już w latach 70. i 80. możemy przywołać liczne wezwania do ograniczenia takich praktyk lub wprowadzenia różnych mechanizmów kontroli nad działaniami państwa, a wkrótce

New York 2018, s. 14–174. Na temat maszynowego spaczenia ocen wykorzystywanych na potrzeby postępowań karnych zob. *A. Renda*, Artificial Intelligence Ethics, governance and policy challenges. Report of CEPS Task Force, Brussel 2019, s. 25–26. Szerzej problemy nieprawidłowości ocen omawia *T. Chivers*, The AI Does Not Hate You. Superinteligence, rationality and the race to save the world, London 2019, s. 143–164.

[3] O tym, dlaczego wyniki w nauce mogą mieć wpływ na stawki ubezpieczeniowe, zob. *V. Mayer-Schoenberger*, *K. Cukier*, Big Data, s. 210.

[4] Sformułowania tego użył np. w tytule XXI Forum Teleinformatyki „Moc danych – nowe źródła i metody analizy i ochrony danych" (Miedzeszyn 24–25 września 2015 r.).

[5] *K. Pries*, *R. Dunningham*, Big Data Analytics. A Practical Guide for Managers, Boca Raton–London–New York 2015, s. 64–66.

również nad działaniami podmiotów rynkowych. W tym znaczeniu możemy powiedzieć, że zastrzeżenia wobec możliwości naruszania wolności i praw informacyjnych lub manipulowania wielkoskalowymi zasobami danych nie są niczym nowym. Tym jednak, co wyróżnia obecną falę zintegrowanego przetwarzania informacji przy wykorzystaniu technologii komunikacyjnych, określanego terminem *Big Data*, jest wszechobecność takich działań i siła.

Liczba urządzeń podłączonych do Internetu przewyższa liczbę ludzi żyjących na Ziemi. Jest to jednak dopiero początek procesu, który zmultiplikuje liczbę urządzeń, dostępną pamięć i pasmo transmisji. Przewiduje się, że „Internet rzeczy" oraz analiza dużych zbiorów danych zostanie dodatkowo wzmocniona przez powiązanie tych działań z systemami opartymi na sztucznej inteligencji[6], przetwarzaniu poleceń i treści zapisanych w języku naturalnym oraz z systemami przetwarzającymi informacje biometryczne (rozpoznającymi głos, wizerunek lub inne indywidualizujące cechy osoby). Choć sama idea zastosowania sztucznej inteligencji dla umożliwienia systemom uczenia się nie jest nowa, dla trzeciej dekady XXI w. będzie to już nie idea, lecz rzeczywistość. Instytucje publiczne i podmioty komercyjne są dziś w stanie wykroczyć poza „eksplorację danych" ku działalności, którą można by nazwać „eksploracją rzeczywistości"[7].

Tak w Polsce, jak i całej Europie panuje przekonanie, że konieczność wykorzystywania takich rozwiązań przez administrację publiczną nie idzie w parze z możliwościami tworzenia i utrzymywania takich rozwiązań przez polskie i europejskie podmioty publiczne. Z pewnością instytucje publiczne – nawet te związane z utrzymaniem bezpieczeństwa publicznego – nie są dziś w stanie samodzielnie prowadzić centrów kompetencyjnych badających i wdrażających prawdziwie innowacyjne metody przetwarzania danych.

Być może jednocześnie jesteśmy już u kresu paradygmatu chmury. Ocenia się, że za kilka lat analiza danych nie będzie przeprowadzana na zasobach, które będą gromadzone w ogromnych centrach przetwarzania danych, jak to dzieje się dziś, z tego powodu, że lawina danych, z którą mamy obecnie do czynienia, spowoduje, że nie będzie takiego miejsca na Ziemi, gdzie te dane będą mogły być przechowywane na stałe[8]. Nie będzie również sensu, aby gromadzić je na bieżąco do zasobu większego niż ten, który potrzebny jest w urządzeniu, na którym są one gromadzone. Tym samym może dojść do swoistego ożywienia i ponownego spopularyzowania modelu gridowego czy też postgridowego. W modelu tym dane przechowywane będą na urządzeniach, analityka będzie zaś dokonywana przy pomocy narzędzi zcentralizowanych, co nie znaczy, że centralnych.

Analiza danych powinna być przeprowadzona po to, żeby osiągnąć cel, którym jest poszerzenie wiedzy podmiotów dokonujących analizy lub podmiotów,

[6] *J. Patterson*, *A. Gibson*, Deep Learning. Praktyczne wprowadzenie, Gliwice 2018, s. 365–374.
[7] *N. Bostrom*, Superinteligencja. Scenariusze, strategie, zagrożenia, Gliwice 2016.
[8] O problemie składowania danych w takich centrach pisze *B. Smith*, Tools and Weapons. The promise and peril of the digital age, London 2019, s. XIII–XXII.

na których rzecz analityk pracuje. Można mieć jednak uzasadnione wątpliwości, czy na pewno jest to „wiedza". Czy na pewno te narzędzia, które mamy, dążą do tego, żeby przekazać nam wiedzę? Można odnieść wrażenie, że tak jak wyrastamy na pewnym etapie naszego życia z wiary w elfy i w Świętego Mikołaja, tak niespecjalnie wyrastamy z wiary w to, że ktoś na zewnątrz posiada wiedzę, którą możemy wykorzystać do naszych celów, pod warunkiem że będziemy mieli odpowiednie narzędzia do jej zdobycia. Chcemy wierzyć, że ktoś lub coś – choćby jakiś system informacyjny – odpowie nam na wielkie pytania, które mu zadamy.

W 2002 r. *D. Kahneman* otrzymał nagrodę Nobla za badania dotyczące tego, kto przekazuje nam wiedzę i na ile ci, którzy wiedzę nam przekazują, są w stanie wpłynąć na to, w jaki sposób ją odczytujemy, nawet jeśli sami głosimy, że nasze rozumienie jest efektem analizy przeprowadzonej na zupełnie obiektywnych danych. W badaniach, które prowadził, potwierdzono, że tak naprawdę autorytet, który przekazuje nam dane, jest znacznie bardziej istotny, niż to co w danych się znajduje. Moja 5-letnia córka wyrasta z okresu, kiedy wierzyła w Świętego Mikołaja. Autorytety – rodzina, książki, filmy – przekazują jej, że Święty Mikołaj istnieje, ale teraz jej „bańka społeczna" w postaci koleżanek i kolegów „sieje zwątpienie", twierdząc, że być może z tym Świętym Mikołajem nie jest do końca tak, jak twierdzą autorytety. Można powiedzieć, że jej analizy przynoszą sprzeczne dane, ale na razie autorytety przeważają. Tymczasem moja starsza córka, przeszedłszy okres wątpliwości kilka lat temu, doszła do punktu, w którym uznała, że bardziej praktyczne jest dalsze powoływanie się na Świętego Mikołaja niż na prawdziwy obraz świata, gdyż wie, że ze Świętym Mikołajem nie trzeba będzie negocjować, podczas gdy rodzice mogą próbować mieć własne zdanie na temat oczekiwanych przez nią prezentów. Natomiast jeżeli Świętemu Mikołajowi przekaże, co chciałaby dostać, to nawet jeżeli w niego nie wierzy, pominąć można nieprzyjemny etap negocjacji z prawdziwymi „decydentami". Można powiedzieć, że ekonomicznie opłacalniejsze jest dla niej podtrzymywanie „starego mitu" przekazanego przez autorytety, nawet wówczas, gdy proponowane niegdyś przez autorytety rozumienie świata okazało się nieprawdziwe lub nieprecyzyjne.

Grupa naukowców, która w 2002 r. pracowała z *D. Kahnemanem*, kilka lat później otrzymała nagrodę *IgNoble* za opracowanie naukowe dotyczące tego, że droga, fałszywa medycyna jest bardziej skuteczna niż tania, fałszywa medycyna. Te badania są chyba jeszcze łatwiejsze do porównania z dzisiejszym podejściem do analizy danych. Można powiedzieć, że jeśli wydaliśmy na dany system informacyjny wykonujący analizę danych odpowiednio duże środki, to „musimy" wierzyć wynikom jego działania. Dochodzimy do punktu, w którym tworzenie rozbudowanych systemów informacyjnych analizujących dane pochodzące z zasobów *Big Data*, jest dla nas samo w sobie powodem do udowadniania, że to co zrobiliśmy ma sens.

Nie jest to nowy fenomen. Tak naprawdę już od lat 70. XX w. mamy do czynienia z analizą wielkoskalowych zasobów danych i na ich przykładzie możemy obserwować opisany wyżej paradoks. Zmienia się jedynie skala przetwarzania danych. Problem podstawowy pozostaje niezmienny. Wciąż musimy zwracać uwagę na to, czy wiemy, jakie dane posiadamy i czy wiemy, na ile te dane reprezentują świat, który chcemy opisać i analizować[9].

Maszynowa analiza danych jest zjawiskiem koniecznym i pozytywnym z punktu widzenia dobra społecznego. Powinniśmy jednak obserwować, co dzieje się z obiektem tej analizy, nie popadając jednocześnie w skrajności w ocenie zalet i wad procesu[10], z drugiej strony nie wprowadzać ideologii do myślenia o tym, w jaki sposób obserwujemy świat i w jaki sposób świat jest dzisiaj zdigitalizowany i zdanetyzowany na co dzień[11].

Innym problemem, z którym stykają się osoby, próbujące poprawić jakość wyników maszynowej analizy danych i jednocześnie chroniąc osoby, których dane są przetwarzane, jest paradoks „minimalizacji danych w zbyt małym zasobie". Jeżeli nie dostarczymy systemowi informacyjnemu wszystkich danych, które są dostępne, będzie on działał na zbiorze zawężonym, niereprezentującym rzeczywistego świata[12]. Tymczasem zasada minimalizacji danych jest wymaganiem RODO[13].

Kolejnym wyzwaniem, którego wpływ na proces przetwarzania automatycznego danych w systemach *Big Data* nie może być niedoceniany, jest zastosowanie sztucznej inteligencji i robotyki do działań dziś wykonywanych wyłącznie przez człowieka. Zaawansowane systemy oparte o rozwiązania sztucznej inteligencji będą oferować rozległy potencjał analityczny, wykraczający poza obecne zastosowania. Twórcy takich systemów dążą do rozwoju automatycznych działań symulujących badania naukowe lub prace analityków. Już teraz algorytmy potrafią rozumieć i tłumaczyć języki, rozpoznawać obrazy, pisać artykuły informacyjne i analizować dane medyczne.

W 2016 r. kancelaria prawnicza *Baker & Hostetler* z centralą w Cleveland w stanie Ohio poinformowała, że w 100-lecie swego istnienia postanowiła jako pierwsza globalna firma prawnicza „zatrudnić" system sztucznej inteligencji Ross jako „adwokata" w swym biurze dla wspierania działu zajmującego

[9] *J. Cheney-Lippold*, We Are Data. Algorithms and the Making of Our Digital Selves, New York 2017, s. 257–264.

[10] Obawy społeczeństwa i pierwsze reakcje rządów w skrótowy sposób opisuje *A. Renda*, Artificial Intelligence, s. 23.

[11] *S. Zuboff*, The Age of Surveillance Capitalism, New York 2018; zob. również *J. Kreft*, Władza algorytmów. U źródeł potęgi Google i Facebooka, Kraków 2019, s. 160–163.

[12] Habermasowskie rozróżnienia systemu i „świata życia" w czasach powszechnej maszynowej analizy danych przeprowadza *E. Finn*, What Algorithms Want? Imagination in the age of computing, Cambridge 2017, s. 108–111.

[13] 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 (ogólne rozporządzenie o ochronie danych) (Dz.Urz. UE L 119, s. 1).

się postępowaniami upadłościowymi. Ross jest systemem opartym na rozwijanych od kilku lat przez IBM rozwiązaniach platformy Watson, oferującej samouczące się i aktywnie korzystające z rozproszonych baz wiedzy narzędzia[14]. Ross został stworzony w celu odczytywania zadawanych w języku naturalnym kwestii, proponowania hipotez odpowiedzi, wyszukiwania i generowania odpowiedzi opatrzonych odwołaniami do literatury i zasobów sieciowych, by wspierać proponowane rozumowanie. Z założenia Ross na bieżąco obserwuje zmiany w prawie (w tym znaczeniu może być uznany za protoplastę nowej linii systemów bezpośrednich, omawianych w kolejnym rozdziale), uzupełniając to informacjami z rozproszonych baz dotyczących orzecznictwa sądowego. Istotną funkcjonalnością systemu jest ograniczanie liczby rezultatów odpowiedzi według wbudowanego, ale samouczącego się algorytmu. Oznacza to, że z tysięcy możliwych odpowiedzi Ross wybierze te, które sam uzna za najważniejsze. O ile w systemach pośrednich decyzje co do wagi i znaczenia odpowiedzi podejmuje redaktor systemu, w przypadku Rossa ta decyzja podejmowana jest przez algorytm wbudowany w program.

Wiedząc, że Watson, będący podstawą techniczną dla Rossa, umożliwia sugerowanie prawidłowej odpowiedzi i nakierowanie samego narzędzia na znalezienie wiarygodnego uzasadnienia stawianej tezy, również w przypadku Rossa pojawiają się wątpliwości co do obiektywności przedstawianej odpowiedzi. Twórcy rozwiązania odpowiadają, że Ross nie „pracuje" jako „sędzia", lecz jako „adwokat" podpowiadający klientowi, jak uzasadniać stanowisko w sprawie. O skuteczności działania Rossa decydować będzie to, na ile przekona on do zdania klienta innych uczestników postępowania. Myśląc o przyszłości, trzeba zastanowić się, na czym polegać będzie rozstrzygnięcie sporu prawnego w sytuacji, gdy narzędzia klasy Rossa doradzać będą wszystkim uczestnikom postępowania z osobna. Ciekawym wątkiem jest również oczekiwanie, że takie systemy bezpośrednie będą uczyły się na własnych sukcesach i błędach w różnych postępowaniach, w których wezmą udział. Kreuje to oczekiwanie, że gdzieś znajdować będzie się moduł porównujący wynik „doradztwa" Rossa wobec różnych klientów. To zaś każe zapytać, czy ów moduł podsumowujący będzie jedynie udoskonalał działanie systemu, czy też w końcu doprowadzi do tworzenia jedynie słusznej linii porad.

Idea zastąpienia prawników przez sprawnie działające systemy bezpośrednie nie jest – jak wspomniano wcześniej – nowa[15], jednak Ross jest z pewnością heroldem zmian, o których pisze *R. Susskind*, przewidując, że rola prawników w przyszłości może być podobna do roli krawców w czasach, gdy

[14] *S. Lohr*, Dataism. Inside the big data Revolution, London 2016, s. 9–10.

[15] *Ł. Goździaszek*, Perspektywy wykorzystania sztucznej inteligencji w postępowaniu sądowym, PS 2015, Nr 10, s. 46–60 oraz *A. Bieliński*, Potencjalne obszary zastosowania sztucznej inteligencji w postępowaniu *cywilnym* – czy obecnie ma to rację bytu i czy jesteśmy na takie rozwiązania gotowi, w: *K. Flaga-Gieruszyńska*, *J. Gołaczyński*, *D. Szostek* (red.), Sztuczna inteligencja, blockchain, cyberbezpieczeństwo oraz dane osobowe. Zagadnienia wybrane, Warszawa 2019, s. 58–61.

większość ubrań wybieramy z oferty sklepowej, a nie szyjemy na miarę. Prawnicy mają, zdaniem *R. Susskinda*[16], wkraczać wówczas, gdy należy dostosować rozwiązanie proponowane przez swoiste supermarkety prawne (systemy bezpośredniego dostępu) do konkretnej sytuacji prawnej (a więc swoiste „poprawki krawieckie"), lub wówczas gdy potrzebujemy rzeczywiście rozwiązania „szytego na miarę" w postaci szczególnej, wyjątkowo istotnej, precedensowej sprawy.

Niewątpliwie celem działań związanych z maszynową analizą danych musi być równolegle poszerzanie możliwości analityki danych i zwiększanie jakości danych. Bardzo istotne jest ocenianie, w jakim celu pierwotnie zbierano dane, gdyż ów cel może znacząco wpływać na kształt zasobu i jego kompletność. Jest to szczególnie ważne, gdy dane takie pochodzą z inteligentnych środowisk[17], w których rozgrywać będą się procesy informacyjne XXI w., takich jak inteligentne miasto czy inteligentny system transportowy. Owe inteligentne środowiska nie są bowiem jednolitymi strukturami stworzonymi przez jednego architekta czy nawet jedną grupę projektową. Nie są nawet zespołem systemów zarządzanych centralnie przez centrum koordynacyjne na poziomie państwa, miasta czy sektora rynku. Podstawowym rozwiązaniem stają się zespoły z zasady otwartych systemów informacyjnych, które umożliwiają dynamiczne dołączanie do architektury stworzonej dla danego środowiska nowych komponentów. Część z systemów pozostaje zamknięta i dostępna jedynie dla głównych twórców danego kompleksu, lecz z założenia należy przyjmować, że i te systemy dążyć będą w najbliższej przyszłości do większej otwartości. Otwartość takich środowisk staje się podstawową wartością sama w sobie. Można zaryzykować stwierdzenie, że jedynie środowiska z zasady otwarte będą mogły skutecznie rozwijać się w sytuacji, gdy sama komunikacja w tych środowiskach opata jest o otwarte sieci. Nie ma wątpliwości, że tworzone będą moduły zamknięte, ale będą one wyjątkiem od generalnej – chronionej przez prawo – zasady otwartości. Otwartość systemów nie będzie wszakże oznaczała całkowitego otwarcia dostępu do wszystkich przetwarzanych zasobów[18].

[16] *R. Susskind*, Koniec świata prawników? Współczesny charakter usług prawniczych, Warszawa 2010 oraz nowsza, nieprzetłumaczona na język polski pozycja tego samego autora dotycząca tego zagadnienia: Tomorrow's Lawyers: An Introduction to Your Future, London 2017.

[17] Inteligentnym środowiskiem nazywamy obszar w świecie fizycznym, który jest bogato nasycony niewidocznymi sensorami, nadajnikami, czytnikami i innymi elementami przetwarzającymi dane wbudowanymi w przedmioty codziennego użytku i stale połączonymi z siecią. Zob. *M. Weiser*, *R. Gold*, *J. S. Brown*, The origins of ubiquitous computing research at PARC in the late 1980s, IBM System Journal 1999, vol. 38 (4), s. 693–696; zob. również *C. Perera i in.*, Context-Aware Dynamic Discovery and Configuration of 'Things' in Smart Environments, w: *N. Bessis*, *C. Dobre* (red.), Big Data and Internet of Things: A Roadmap for Smart Environments, Springer 2014, s. 215–218.

[18] *E. Morozov*, To Save Everything Click Here, London 2013, s. 63–99.

Streszczenie

Skupiając się na rozwoju narzędzi analitycznych i możliwościach ich użycia nie powinno się zapominać, że podstawą dla prawidłowej oceny świata poddawanego analizie jest poprawne określenie, jakie dane i dla jakiego celu chcemy przetwarzać. Analiza danych powinna być przeprowadzona po to, żeby osiągnąć cel, którym jest poszerzenie wiedzy podmiotów dokonujących analizy lub podmiotów, na których rzecz analityk pracuje. Celem działań związanych z maszynową analizą danych musi być równolegle poszerzanie możliwości analityki danych i zwiększanie jakości danych. Istotne jest ocenianie, w jakim celu pierwotnie zbierano dane, szczególnie, gdy dane takie pochodzą z inteligentnych środowisk, w których rozgrywają się procesy informacyjne XXI w., takich jak inteligentne miasto czy inteligentny system transportowy. Ten pierwotny cel może znacząco wpływać na kształt zasobu i jego kompletność. Instytucje publiczne i podmioty komercyjne są już w stanie wykroczyć poza „eksplorację danych" ku działalności, którą można by nazwać „eksploracją rzeczywistości".

Abstract

Focusing on the development of analytical tools and the possibilities of their use, it should not be forgotten that the basis for the correct assessment of the world being analyzed is the correct determination of what data and for what purpose we want to process. Data analysis should be carried out in order to achieve the goal of expanding the knowledge of entities performing the analysis or entities for which the analyst works. The goal of machine data analysis activities must be to simultaneously expand data analytics capabilities and increase data quality. It is important to assess the purpose for which data was originally collected, especially when such data comes from intelligent environments in which information processes of the 21st century take place, such as a smart city or intelligent transport system. This original goal can significantly affect the shape of the resource and its completeness. Public institutions and commercial entities are already able to go beyond „data mining" towards activities that could be called „reality mining".

Marek Niezgódka

Rozdział 2.
e-Infrastruktury analityki danych: wybrane problemy strukturalne

1. Uwagi wprowadzające

Dziedzina analityki danych obejmuje procesy analizy dowolnych rodzajów danych w celu wydobycia z nich informacji, realizowane z wykorzystaniem specjalizowanych systemów przetwarzania i oprogramowania. W obecnym opracowaniu zwrócimy uwagę na wybrane aspekty wpływu e-infrastruktur analityki danych na wartość wyników analiz uzyskiwanych z ich użyciem.

Jest to specyficzna perspektywa, która skupia się w pierwszym rzędzie na wewnętrznych uwarunkowaniach związanych ze strukturą systemów analityki, obejmującą zarówno same techniczne infrastruktury informatyczne, jak też realizowane przez nie serwisy (Infrastruktura jako serwis), a nie na relacjach takich systemów z ich zewnętrznym środowiskiem.

Odnotujmy jedynie, że szeroko rozumiana problematyka bezpieczeństwa w Internecie jest głównie traktowana z perspektywy zagrożeń naruszania integralności i zawartości treściowej różnych zasobów lub niepożądanej ingerencji w takie zasoby, połączonej z naruszaniem prywatności. Z przyjętej przez nas perspektywy właściwością o szczególnym znaczeniu w kontekście bezpieczeństwa pełnego cyklu gospodarki danymi jest uwarunkowana strukturalnie odporność (niska wrażliwość na szeroki zakres zaburzeń) obsługujących ją e-infrastruktur.

W odniesieniu do analityki danych takie podejście oznacza, że głównym obiektem zainteresowania będzie nie tyle sama zewnętrzna ingerencja w system ani wywoływane przez nią potencjalne naruszenie lub nawet destrukcja zasobów danych na dowolnym etapie cyklu ich przetwarzania: od akwizycji i kuracji, przez strukturyzację, do agregacji i archiwizacji, co wynikający z cech strukturalnych e-infrastruktury wpływ na zaburzenia realizacji procesów analizy takich danych i wyników generowanych przez te procesy.

Jest to bardzo szeroki zakres problematyki, nawet jeśli ograniczyć się do jej bardzo ogólnego przedstawienia. Dlatego nasza prezentacja będzie odnosić się do ograniczonego obszaru zastosowań analityki danych dla ważnego zakresu zastosowań związanego z inteligentnymi systemami energetycznymi, szczególnie istotnego ze względu na znaczący udział w nich odnawialnych technologii generacji energii, aktualnych ze względu na ich strategiczne znaczenie gospodarcze, a zarazem reprezentatywnych dla szerokiej klasy złożonych inteligentnych systemów sieciowych.

Jeszcze raz podkreślamy, że nie będziemy zajmować się bezpośrednio problemami wiarygodności wyników analiz danych w kontekście celowych destrukcyjnych działań stron trzecich, co stanowi wielki, a zarazem krytyczny obszar zagrożeń wartości wszelkich procesów realizowanych na danych. Natomiast same infrastruktury informatyczne używane w pełnym cyklu takich procesów są źródłem szeregu często znaczących problemów, którym poświęcimy tu uwagę. Tylko część tych problemów jest pochodną różnego rodzaju niepożądanych ingerencji w funkcjonowanie infrastruktur, inne mogą wynikać z powodu zaburzeń technicznych, w każdym przypadku naruszona może zostać wiarygodność generowanych wyników przetwarzania i analityki.

Taki wybór odniesień aplikacyjnych wynika również z realizowanego w Centrum Cyfrowej Nauki i Technologii UKSW szerokiego zakresu projektów dotyczących problematyki analityki danych dla inteligentnych systemów sieciowych, w pierwszym rzędzie wywodzących się z energetyki z istotnym udziałem generacji odnawialnej (w tym z realizowanego w ramach Kontraktu Terytorialnego dla Mazowsza projektu utworzenia Multidyscyplinarnego Centrum Badawczego i jego kampusu).

2. E-Infrastruktury analityki danych w inteligentnych systemach sieciowych

Rozszerzona analityka danych (*augmented data analytics*) obejmuje zastosowania metod uczenia maszynowego i przetwarzania w języku naturalnym dla udoskonalenia standardowej (opisowej) analityki danych, ich udostępniania oraz analityki biznesowej. Jej specyficznymi obszarami są analityka predykcyjna i preskrypcyjna[1].

Jedynie część danych może być zbierana z użyciem inteligentnych sensorów, taka możliwość nie obejmuje w szczególności zdecydowanej większości danych społecznych.

[1] Zob. The Age of Analytics: Competing in a Data-Driven World, McKinsey Global Institute, December 2016; Top 10 Strategic Technology Trends for 2019: A Gartner Trend Insight Report, March 2019; Gartner Identifies Top 10 Data and Analytics Technology Trends for 2019, February 18, 2019, Sydney; Big Data Is Dead. Long Live Big Data AI. In: Forbes, July 1, 2019.

Z tego powodu krytyczne znaczenie dla wielu obszarów zastosowań ma odporność (*robustness*) e-infrastruktury systemów na jak najszerszą grupę zaburzeń.

Dla jednoznaczności przyjmiemy, że przez e-infrastrukturę systemu będziemy rozumieć sumę zintegrowanych zasobów technicznych technologii cyfrowych, obliczeniowych i komunikacyjnych, a także serwisów zapewniających dostęp do wszelkich rozproszonych zasobów cyfrowych w tych systemach.

W kontekście analityki danych można wprowadzić rozróżnienie e-infrastruktur ze względu na tryb przetwarzania na systemy:

1) analityki czasu rzeczywistego – dla danych zbieranych *online*;
2) analityki *offline* – w sytuacjach, kiedy czas do uzyskania wyników analiz nie jest krytyczny;
3) analityki na poziomie pamięci – kiedy wielkość zbiorów danych używanych do analizy umożliwia ich bieżące umieszczenie w pamięci operacyjnej;
4) analityki masowej – kiedy wielkość analizowanych zbiorów przekracza rozmiar zasobów bezpośrednio dostępnych w systemach obliczeniowych i wymaga użycia rozproszonych modeli przechowywania danych i ich przetwarzania.

Naturalną pochodną powyższego rozróżnienia e-infrastruktur analityki danych jest uwzględnienie odrębności[2]:

1) modeli przetwarzania rozproszonego sieciowo, co obejmuje architektury gridowe oraz chmurowe, działające w trybie zdalnym;
2) modelu przetwarzania skupionego, zakładającego dostępność systemów obliczeniowych (klastrów i innych klas systemów wieloprocesorowych) bezpośrednio w ośrodkach gromadzenia i przechowywania danych.

W zależności od przyjętego modelu przetwarzania różna jest proporcja czynników generujących możliwe zaburzenia wyników analizy danych; inne są wzajemne relacje wagi warstwy komunikacyjnej i warstwy przetwarzania obliczeniowego, inna jest natura wrażliwości całościowego systemu.

Problematyka oceny wrażliwości stanowi obszar krytycznych wyzwań dla analityki wielkich danych, szczególnie w odniesieniu do e-infrastruktur systemów o wysokiej skali złożoności przestrzennej i wielu skalach dynamiki.

Jest tak w szczególnie skrajnej formie dla systemów energetycznych, gdzie zarówno relacje przestrzenne, jak i znaczenie ekstremalnie szybkich procesów odgrywają centralną rolę. Wyzwaniem jest tam harmonizacja akwizycji danych od poziomu pojedynczych sensorów do instalacji technologicznych oraz ich transferu do centrów przetwarzania z procedurami przetwarzania, obejmującymi pełny cykl od strukturyzacji danych, przez warstwę obliczeniową, do postprocessingu. Każdy z tych etapów charakteryzuje się specyficznymi cecha-

[2] Zob. *P. Raj i in.*, High-Performance Big-Data Analytics: Computing Systems and Approaches, Computer Communications and Networks, Springer 2015.

mi wrażliwości, co odbija się na odporności systemu jako całości, a tym samym na wiarygodności i użyteczności wyników analityki danych.

W przypadku e-infrastruktury sieciowych inteligentnych systemów energetycznych wzajemne oddziaływania infrastruktury technicznej, procesów komunikacji oraz warstwy obliczeniowej analityki odgrywają istotną rolę ze względu na wyjątkowo silne nagromadzenie wszystkich wyżej wymienionych efektów.

Czynnikiem o szczególnie istotnym wpływie na wrażliwość procesów analityki danych dla takich systemów jest ich przestrzennie rozproszony układ. Przede wszystkim rozproszenie przestrzenne wynika z decentralizacji i delokalizacji systemów przetwarzania, zarówno na poziomie archiwizacji danych, jak i organizacji procesów obliczeniowych, dodatkowo jednak nawet lokalnie w centrach przetwarzania wiąże się ono z wieloprocesorową architekturą samych systemów komputerowych. Rozproszone modele mogą ponadto generować problemy skalowalności tak w odniesieniu do sprawności obsługi wielkich zbiorów danych, jak i zrównoważonej realizacji obliczeń, a tym samym powodować zaburzenia synchronizacji procesów przetwarzania.

W przypadku szybkich danych, wymagających uzyskiwania wyników ich analizy w skrajnie krótkim czasie, problem skalowalności staje się krytyczny: zarówno opóźnienia komunikacyjne, jak i skalowalność obliczeń są czynnikami generującymi ryzyko utraty wartości przetworzonych danych wskutek ekscesywnego czasu do wyniku[3].

Zagrożenia te są szczególnie mocno eksponowane w przypadku standardowego modelu chmurowego (*cloud computing*), w którym zapewnienie uzyskania wyników w ekstremalnie krótkim czasie jest często problematyczne[4]. Drogą prowadzącą do gwarantowanego uzyskania wyników w narzuconym czasie jest użycie jakiegoś wariantu wielopoziomowej hierarchii zmodyfikowanych modeli typu chmurowego, takich jak modele przetwarzania mgłowego (*fog computing*)[5] czy związany z nimi model przetwarzania krawędziowego (*edge computing*)[6]. Wprowadzane w tych modelach bezpieczne mechanizmy transferu danych zapewniają nie tylko podniesienie poziomu bezpieczeństwa danych, ale także obniżenie w wielu wymiarach wrażliwości całego systemu. Równocześnie, przyjęcie hierarchicznych struktur przetwarzania radykalnie ogranicza wielkość transferów danych w całym systemie, zapewniając wyższą

[3] Zob. *P. Raj i in.*, High-Performance Big-Data Analytics.
[4] Zob. NIST Special Publication 800-145: The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology, September 2011.
[5] Zob. NIST Special Publication 500-325: Fog Computing Conceptual Model: Recommendations of the NIST, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.500-325.pdf (dostęp: 15.9.2019 r.).
[6] Zob. *Ch. Mahmoudi*, *F. Mourlin*, *A. Battou*, Formal Definition of Edge Computing: An Emphasis on Mobile Cloud and IoT Composition, https://ws680.nist.gov/publication/get_pdf.cfm?pub_id=919490 (dostęp: 15.7.2019 r.).

skalowalność procesów przetwarzania, przez co umożliwia uzyskanie radykalnie krótszego czasu do osiągnięcia wyniku analizy[7].

3. Analityka danych w inteligentnych systemach sieciowych, z odniesieniami do systemów energetycznych

Inteligentne systemy energetyczne, traktowane jako charakterystyczna klasa inteligentnych systemów sieciowych, łączą szeroki zakres składowych związanych z produkcją i konsumpcją całego spektrum form energii z warstwą e-infrastruktur, realizujących pełny zakres procesów gospodarki danymi dla optymalizacji podejmowania decyzji, zarządzania i sterowania, aż do poziomu analityki danych w czasie zbliżonym do rzeczywistego[8]. Wymagania funkcjonalne wobec takich infrastruktur, dotyczące realizowanych przez nie procesów analizy danych, obejmują:

1) niskie opóźnienia (*low latency*);
2) wysoką sprawność przetwarzania (*high performance*);
3) nieprzerwaną dostępność serwisów;
4) wysoki poziom bezpieczeństwa pełnego cyklu przetwarzania;
5) elastyczność i skalowalność, od poziomu przetwarzania wymaganego dla wielkich danych, aż do trybu czasu rzeczywistego.

Dla inteligentnych systemów energetycznych zagadnienia analityki danych mają pod wieloma względami charakter wielkoskalowy:

1) wielkość zbiorów danych tak duża, że przekracza typowe lokalne możliwości archiwizacyjne: jednym ze sposobów pokonywania takich ograniczeń jest rozproszona archiwizacja danych surowych, w szczególności pochodzących z pomiarów u końcowych użytkowników systemu oraz wszelkiego rodzaju działających w nim innych układów pomiarowych. Wielkość tych zbiorów wynika z liczby punktów akwizycji danych u końcowych odbiorców energii, którzy mogą być równocześnie jej producentami, jak również ze złożoności i liczby innych istotnych źródeł danych, w tym różnorodnych układów sensorycznych;
2) szybkość i intensywność zbierania oraz sieciowego transferu danych, w szczególności sygnałowych, bliska wymianie danych w czasie rzeczywistym: przy tym liczba punktów akwizycji danych będzie docelowo co najmniej porównywalna z wielkością populacji użytkowników systemu. W tym przypadku dane są określane jako „szybkie" (*fast data*), co narzuca

[7] Zob. *A. Luntovsky*, *J. Spillner*, Smart Grid, Internet of Things and Fog Computing, In: Architectural Transformations in Network Services and Distributed Systems, Springer Nature 2019, s. 135–210; *F.Y. Okay*, *S. Ozdemir*, A fog computing based smart grid model, In: IEEE 2016 International Symposium on Networks, Computers and Communications (ISNCC), Hammamet, IEEE Xplore Digital Library, DOI: 10.1109/ISNCC.2016.7746062.

[8] Zob. Intelligent Energy Systems: A White Paper with Danish perspectives, http://www.ea-energianalyse.dk/reports/901_white_paper_intelligent_energy_systems_2010.pdf (dostęp: 15.9.2019 r.).

ekstremalne wymagania na sprawność urządzeń pomiarowych, efektywną sprawność szerokopasmowych sieci telekomunikacyjnych, wykorzystujących technologie, takie jak LTE 450 i 5G. Co więcej, efektywność podejmowanych decyzji narzuca równie ostre wymagania na tryb realizacji procesów przetwarzania tych danych i ich analityki. Niedotrzymanie rygorów czasowych prowadzi w skrajnych sytuacjach do utraty wartości użytkowej wyników analityki danych;

3) różnorodność rodzajów danych, jedynie w małej części systematycznie ustrukturyzowanych: w zdecydowanej większości procesów konieczna jest akwizycja danych nieustrukturyzowanych, takich jak komunikaty, formy komunikacji mediów społecznych, zbiory multimodalne (obrazy cyfrowe, ciągi sygnałowe z sensorów, nagrania wideo-audio);
4) gęstość istotnej informacji zawartej w danych: miarą inteligencji procesów akwizycji danych jest zachowanie w nich wysokiej wartości użytecznej informacji, wydobywanej w procesach analizy, niezależnie od wielkości zbiorów danych objętych przetwarzaniem.

Od e-infrastruktur analityki danych w systemach energetycznych, obok zdolności do spełniania powyższych warunków, wymagana jest strukturalna odporność na szeroki zakres zaburzeń operacyjnych. Przestrzennie rozproszona natura tych infrastruktur obejmuje nie tylko systemy transmisji danych, ich przechowywania, ale – co szczególnie krytyczne – przetwarzania na wszystkich etapach strukturyzacji, analityki i bazującego na danych modelowania obliczeniowego.

Jako pozornie racjonalny wybór modelu operacyjnego e-infrastruktury może narzucać się architektura chmury[9]. Rozproszone architektury w standardowym modelu chmurowym mają jednak dla inteligentnych systemów energetycznych jedynie ograniczony zakres przydatności. Wiąże się to z całym szeregiem zagrożeń gwarancji jakościowych serwisów, których wyeliminowanie wymaga wprowadzania kosztownych procedur, takich jak:

1) redundancja serwisów, w realizacjach wielolokalizacyjnych z automatycznym przełączaniem, tyle że łączy się to z nieuchronnym ryzykiem naruszania ograniczeń czasowych;
2) wprowadzanie dedykowanych, na zasadzie wyłączności, kanałów transmisji;
3) wprowadzania w kontraktach serwisowych skrajnie wymagających klauzul co do SLA (*Service Level Agreement*).

Warto w tym miejscu podkreślić, że wszystkie wyliczone opcje rozwiązań zabezpieczających, które podnoszą poziom odporności na różnego rodzaju zakłócenia operacyjne, nie eliminują zagrożenia utratą jakości serwisów, a jedynie chronią przed naruszaniem ich ciągłej dostępności. Utrzymanie pełnej

[9] Zob. NIST Special Publication 800-145: The NIST Definition of Cloud Computing.