ISSN 2080-1807

TORUŃSKIE STUDIA BIBLIOLOGICZNE
2016, nr 2 (17)

Seminarium „Big Data i cloud computing jako nowe narzędzia w informacji i nauce"

DOI: http://dx.doi.org/10.12775/TSB.2016.030

Istotą wielkich zbiorów danych (ang. *big data*) jest przetwarzanie różnych typów danych równocześnie, co nie jest możliwe w przypadku tradycyjnych systemów. Działania takie obejmują gromadzenie, przechowywanie, wyszukiwanie, analizę i wizualizację danych. Charakteryzując zbiory typu big data, mówi się o trzech wymiarach (tzw. 3V):

- *Volume* (ilość): wielkie dane (rozpoczynają się od zbiorów petabajtowych);
- *Variety* (różnorodność): odnosi się do wielu różnych typów danych i plików, dla których tradycyjne relacyjne bazy danych nie są dostosowane (pliki dźwiękowe i wideo, dokumenty, dane geolokacyjne, logowania sieciowe, linki tekstowe);
- *Velocity* (prędkość): szybkość aktualizacji i używania danych[1].

Tak więc big data opiera się na analizie dużej ilości różnorodnych danych w czasie zbliżonym do rzeczywistego (wielkość, szybkość, odmiany) i zazwyczaj polega na zastosowaniu technologii NoSQL oraz rozproszonej architektury do analizy danych. Analiza przeprowadzana jest w chmurze infrastruktury publicznej, prywatnej lub hybrydowej. Chmura obliczeniowa (ang. *cloud computing*) odnosi się do stosowania skalowalnej sieci komputerowej do wykonywania zadań obliczeniowych, jak Amazon Web Services, Azure czy Google w chmurze[2].

[1] P. Płoszajski, *Big Data. Nowe źródło przewag i wzrostu firm.* „E-mentor" [online] 2013, nr 3 (50) [dostęp 31 lipca 2016]. Dostępny w World Wide Web: http://www.e-mentor.edu.pl/artykul/index/numer/50/id/1016.

[2] M. Ambrusti in., *A view of cloud computing*. „Communications of the ACM CACM" [online] 2010, nr 4 (53) [dostęp 10 maja 2016]. Dostępny w World Wide Web: http://dl.acm.org/citation.cfm?id=1721672.

Poznaniem i omówieniem tych zjawisk zajęli się licznie zgromadzeni uczestnicy seminarium „Big Data i cloud computing jako nowe narzędzia w informacji i w nauce", które odbyło się 9 marca 2016 r. w gmachu Biblioteki Uniwersytetu Warszawskiego (dalej: UW). Zorganizowane zostało ono przez Centrum Europejskie Uniwersytetu Warszawskiego przy wsparciu firmy Microsoft, Interdyscyplinarnego Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego oraz Stowarzyszenia Bibliotekarzy Polskich (dalej: SBP). To ogólnopolskie spotkanie naukowców, bibliotekarzy i specjalistów branży IT otworzyli prof. Dariusz Milczarek, prof. Marta Grabowska (oboje reprezentujący Centrum Europejskie UW), Rafał Czupryński (Microsoft) oraz Elżbieta Stefańczyk (SBP). Seminarium składało się z trzech sesji, w ramach których zaprezentowano siedem wystąpień.

Moderatorem sesji pierwszej, dotyczącej teoretycznych zagadnień informatycznych, był dr Mikołaj Rakusa-Suszczewski. W świat technologii informatycznych i w zagadnienia związane z chmurą obliczeniową wprowadził słuchaczy Karol Żak, entuzjasta najnowszych technologii, reprezentujący firmę Microsoft. Przewodnim tematem jego wystąpienia było omówienie nowego podejścia w branży technologii informacyjnej, zapewniającego firmom i instytucjom dostęp do różnorodnych programów i usług, a pozwalającego osiągać „więcej i szybciej" dzięki możliwości wykorzystania ogromnych centrów danych i usług IT. Karol Żak podkreślił fakt pominięcia obligatoryjności samodzielnego ich tworzenia, zarządzania nimi oraz ich obsługiwania. Prelegent dokładnie omówił kwestię Microsoft Azure – otwartej i elastycznej platformy przetwarzania w chmurze, umożliwiającej szybkie tworzenie i wdrażanie aplikacji oraz zarządzanie nimi w globalnej sieci centrów danych kierowanych przez firmę Microsoft. Wyjaśnił, że używany jest tutaj termin *wirtualny serwer*, ponieważ tak naprawdę nie ma potrzeby wynajmowania sprzętu. Wirtualizacja pozwala na uruchamianie wielu systemów operacyjnych na jednej maszynie oraz ułatwia testowanie nowego oprogramowania i lepiej wykorzystuje posiadane zasoby sprzętowe. To właśnie dzięki wirtualizacji można wykorzystywać przez Internet tę samą maszynę wiele razy. Oczywiście chmura nie ogranicza się do jednego serwera, ale do ich większej puli tworzących *data center.* Infrastruktura chmury składa się z rzeczywistego sprzętu, który jest uzupełniony o mechanizm dostawy. Główną różnicą w stosunku do dotychczasowych modeli jest to,

że dane są oderwane od konkretnego miejsca. W chmurze może funkcjonować platforma taka jak Windows Azure, która zapewnia prefabrykaty do budowanych własnych aplikacji. Użytkownik dysponuje gotowymi funkcjami, interfejsem użytkownika, mechanizmami do zarządzania i administracji, a także sprawdzone modele bezpieczeństwa. Ponadto zyskuje niższe koszty niż w przypadku posiadania serwerowni w firmie (ang. *on premise*), bezpieczeństwo, gdyż w chmurach najczęściej wprowadza się już sprawdzone modele bezpieczeństwa oraz szybkie prototypowanie – czyli szybkie tworzenie i wdrażanie koncepcji aplikacji bez pisania kodu.

Po tym wystąpieniu odbyła się krótka dyskusja, której głównymi tematami były koszty wdrażania i przenoszenia serwerów oraz ich obsługi, a także zagadnienia związane z przenoszeniem danych po wygaśnięciu umowy z firmą je przechowującą.

W sesji drugiej – prowadzonej przez prof. Martę Grabowską – pierwszy referat „Nowoczesne metody analizy danych w chmurze obliczeniowej Microsoft" zaprezentował Radosław Łebkowski, architekt rozwiązań Business Intelligence Microsoft. We wstępie prelegent wskazał na wyzwania w łatwym dostępie do informacji, a mianowicie brak wiedzy i umiejętności użytkowników, konieczność integracji z istniejącymi narzędziami, bezpieczeństwo i zarządzanie danymi oraz rosnąca złożoność danych i źródeł. Wykład tego praktyka pozwolił zgromadzonemu audytorium na poznanie metody realizacji projektów big data opartych na usługach Microsoft Azure (Azure Data Factory, Hadoop on Azure, bazy danych i hurtownie danych w chmurze, pakiet Cortana Analytics Suite). Referent przedstawił zaawansowane metody analizy w oparciu o język R (in-database analytics w SQL Server R Services), a także modelowanie i wizualizację danych wykorzystujących takie narzędzia, jak Microsoft Excel, Power BI, Microsoft SQL Server, Datazen – pozwalające na szybkie przetwarzanie ogromnych zbiorów danych (o rozmiarach petabajtów). Referent dużo miejsca poświęcił pakietowi do obsługi danych big data i zaawansowanej analityki, który umożliwia przekształcanie danych w inteligentne działania – Cortana Analytics Suite. Wskazał na jego zalety, tj. możliwość rozpoczęcia efektywnej pracy poprzez bazowanie na specjalistycznych rozwiązaniach branżowych albo dostosowywanie do swoich potrzeb modeli uczenia maszynowego, interfejsów API i szablonów dostępnych w galerii rozwiązań. Podkreślił również możliwość

wykorzystywania wszystkich danych (dostęp do danych o każdej wielkości, typie i szybkości, działania lokalnie i w chmurze oraz maksymalne ich wykorzystywanie). Wskazał także na otwartość i rozszerzalność, ponieważ użytkownicy mogą pracować ze wszystkimi językami i środowiskami, które już dobrze znają i którymi się posługują, jak np. R, Python i Hadoop. Na koniec prelegent omówił obszar zastosowania Cortana Analytics Gallery w edukacji, analizę danych nieustrukturyzowanych, analizę obrazów czy generowanie miniatur. Zachęcał także do korzystania z darmowej aplikacji Power BI, demonstrując możliwości umieszczania danych naukowych na blogach. W sposób bardzo interesujący pokazał, w jaki sposób wyniki na zadawanie pytania w języku naturalnym (analiza tekstu) mogą być generowane w postaci raportu[3].

Następnie głos zabrała Lidia Stepińska-Ustasiak reprezentująca Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego UW (dalej: ICM). W referacie zatytułowanym „Ocean możliwości – nowe narzędzia w analizie danych" prelegentka starała się odpowiedzieć na pytania, dlaczego treści cyfrowe są ważne dla rozwoju nauk oraz jakie rozwiązania i infrastruktura wdrażane są na potrzeby nauki, a także zademonstrować przykłady interdyscyplinarnych projektów naukowych. W pierwszej części referatu zostały omówione cztery paradygmaty zmian w nauce (ang. *the fourth paradygmat*): empiryczny, teoretyczny, obliczeniowy i eksploracja danych (obecnie trwający) oraz przykłady finansowania infrastruktury repozytoryjnej (projekty Driver, Driver II), a także tworzenia konsorcjów naukowych typu DARIAH.EU, DARIAH.PL. W drugiej części wystąpienia Stepińska-Ustasiak przybliżyła platformę Infona tworzoną przez ICM, będącą portalem komunikacji naukowej, który umożliwia m.in. prowadzenie własnych badań naukowych w obszarze nauk obliczeniowych i modelowania złożonych układów i procesów; tworzenie i rozwój infrastruktury informatycznej rozległych systemów zasobów wiedzy, obejmujących Wirtualną Bibliotekę Nauki, repozytoria zasobów publikacyjnych i bibliograficznych, polskich i zagranicznych, archiwa danych naukowych; podejmowanie działań na rzecz otwartego dostępu do wiedzy poprzez zaangażowanie w projekty ukierunkowane na tworzenie i udostępnianie sieciowych zasobów dla nauki, edukacji

[3] *Bring your data to life with Microsoft Power BI* [online] [dostęp 31 lipca 2016]. Dostępny w World Wide Web: https://powerbi.microsoft.com/en-us/.

i otwartego społeczeństwa wiedzy (projekty Synat, Polon, OpenAire, OpenAirePlus, EuDML, Otwórz Książkę, Platon i inne). W tej części wystąpienia prelegentka scharakteryzowała również narzędzia służące wizualizacji danych naukowych, jak Visnow – laboratorium Analizy Wizualnej, CERMINE (wydobywanie metadanych z plików PDF), Applied Data Analysis Lab-ADALab (m.in. analizy big data z wykorzystaniem systemów Hadoop i Apache Spark). Na zakończenie opowiedziała o współpracy w projektach interdyscyplinarnych, np. z Instytutem Kardiologii w Aninie (pierwszy zabieg operacji tętnicy wieńcowej przy zastosowaniu okularów Google ze zintegrowanym 3D obrazem tomograficznym).

W czasie dyskusji padły pytania dotyczące analizy językowej, m.in. czy narzędzia potrafią przeanalizować emocje oraz niuanse kulturowe. Stepińska-Ustasiak odpowiedziała, iż ICM pracuje nad algorytmem potrafiącym uchwycić siedem podstawowych emocji. Przypomniano, że w Polsce istnieją firmy specjalizujące się w tworzeniu algorytmów językowych.

W sesji trzeciej – moderowanej przez Elżbietę Stefańczyk – dotyczącej zastosowania zbiorów big data oraz realizacji międzynarodowych projektów inicjowanych przez biblioteki i ośrodki badawcze jako pierwsza wystąpiła prof. Marta Grabowska z Centrum Europejskiego UW. W swoim wykładzie referentka przedstawiła rozważania na temat zbiorów big data w bibliotekach i ośrodkach informacji. Scharakteryzowała konsorcja infrastruktury badawczej, jak ERIC, CLARIN i DARIAH.EU, umożliwiające badaczom z nauk humanistycznych pracę z bardzo dużymi zbiorami tekstów. Prelegentka podzieliła zbiory big data na ustrukturyzowane (proste bazy relacyjne, hurtownie danych) oraz nieustrukturyzowane (NoSQL, NLP). Następnie szczegółowo omówiła zadania big data, jakimi są pozyskiwanie, gromadzenie i organizowanie danych, analiza danych (narzędzia do analizy, monitoring drzewa klasyfikacyjnego, sieci neuronowe), raportowanie i wizualizacja. Podkreśliła przy tym ważkość kompozycji architektonicznej całego systemu (platformy, middleware, API). Grabowska zdefiniowała, czym są zbiory big data w rozumieniu bibliotek (zawartość minimum 1 mln zdigitalizowanych dokumentów o objętości powyżej 200 stron każdy). Uzasadniała, że w tym kontekście „Projekt Gutenberg" nie jest zasobem big data. Jest nim z kolei projekt finansowany z Fundacji Mellona, realizowany przez Uniwersytety Indiana i Illinois, zawierający ok. 10 mln tomów, 5 mln

tytułów (w sumie ponad 3 mln stron, ogółem 473 TB) czy system IBM Watson będący rozwiązaniem dla medycyny zawierającym ponad 1 mln książek.

Jako kolejny głos zabrał dr Maciej Szablewski, kierownik Zakładu Zbiorów Bibliologicznych Biblioteki Narodowej. W wystąpieniu zatytułowanym „Humanistyka bez bibliotek jest martwa. Projekt DARIAH" omówił założenia projektu DARIAH.PL i zagadnienia współpracy polskich instytucji naukowych w zakresie popularyzacji humanistyki cyfrowej. Prelegent przybliżył strukturę i kompetencje konsorcjum, w skład którego wchodzi 18 wiodących w dziedzinie humanistyki cyfrowej instytucji naukowych, m.in. Uniwersytet Warszawski jako koordynator i Biblioteka Narodowa.

Ostatnie dwa wykłady zamykające seminarium zostały poświęcone omówieniu założeń „Polskiej Bibliografii Literackiej" [dalej: PBL] jako narzędzia badawczego Instytutu Badań Literackich Polskiej Akademii Nauk [dalej: IBL PAN] oraz zaprezentowaniu konsorcjum CLARIN.PL – polskiej części ogólnoeuropejskiej infrastruktury naukowej, obejmującej cyfrowe archiwa tekstów, danych tekstowych oraz narzędzia do ich automatycznej analizy. Zastępca dyrektora IBL PAN, dr Maciej Maryl, w wystąpieniu „Polska Bibliografia Literacka jako narzędzie badawcze. Założenia teoretyczne" (współautor Piotr Wciślik) omówił remediację bibliografii, wyzwania i wyjście poza prostą remediację, tj. nowe typy dokumentów (dyskurs elektroniczny, dokumenty cyfrowe) czy powiązania z repozytoriami. Prelegent przedstawił projekt „Komputerowa baza danych PBL online", który jest kontynuacją drukowanej bibliografii, opartej na koncepcji bibliografii literackiej prof. Stefana Vrtela-Wierczyńskiego, a będącej podstawowym źródłem informacji bibliograficznej dla badaczy literatury, nauczycieli i studentów filologii, teatrologów, filmoznawców, dziennikarzy, socjologów itd. Transformacja PBL polegała na odzwierciedleniu struktury bibliografii drukowanej w formie elektronicznej bazy danych, utworzeniu formularzy do wprowadzania informacji bibliograficznych oraz licznych indeksów i kartotek obsługujących bibliografię. Dzięki współpracy z Instytutem Informatyki Politechniki Poznańskiej opracowano szczegółową metodologię wprowadzania zapisów źródłowych, elektroniczne sposoby ich organizacji i udostępniania w oparciu o serwer baz danych Oracle. Dalszy rozwój PBL online pozwoli użytkownikom na dotarcie nie tylko do zapisów bibliograficznych, lecz

także do pełnych tekstów publikacji, plików graficznych ze skanami, plików dźwiękowych z zapisem audycji radiowej lub wywiadu, plików wideo z życia literackiego i teatralnego itp. W najbliższej przyszłości PBL online zostanie połączona z zasobami polskich bibliotek cyfrowych.

Ostatnim prelegentem był dr Maciej Piasecki, reprezentujący Zakład Sztucznej Inteligencji Politechniki Wrocławskiej, z referatem „CLARIN. PL – narzędzie badawcze wspierające analizę tekstowych Big Data". CLARIN.PL to polskie konsorcjum naukowe, część ogólnoeuropejskiej infrastruktury badawczej CLARIN. Tworzy je sześć jednostek naukowych, w których powstają elektroniczne zasoby językowe i narzędzia do pracy z dużymi zbiorami tekstów w języku polskim. Polski węzeł CLARIN – Centrum Technologii Językowych – powstaje na Politechnice Wrocławskiej. Dzięki ścisłemu przestrzeganiu przyjętych standardów zarejestrowani użytkownicy uzyskają za pośrednictwem Centrum CLARIN. PL bezpłatny dostęp do narzędzi i zasobów językowych udostępnionych zarówno w Polsce, jak i w centrach CLARIN pozostałych państw członkowskich. Zadania Centrum obejmują również budowę repozytorium, w którym zgromadzone narzędzia i zasoby oznaczane są trwałymi identyfikatorami. Następnym obowiązkiem jest dbałość o techniczną spójność powstającego systemu oraz o przestrzeganie przyjętych standardów, praw dotyczących własności intelektualnej, licencji i zasad etycznych. Centrum ma również za zadanie ustanowienie polityki bezpieczeństwa, np. poprzez certyfikację serwerów i odpowiedzialne zarządzanie danymi osobowymi. Prelegent zapoznał uczestników seminarium z programami, które powstają w Centrum Technologii Językowych konsorcjum CLARIN, a pomocne są m.in. w ustalaniu autorstwa tekstów anonimowych, określaniu profilu psychologicznego autora, automatycznym streszczaniu, wydobywaniu z tekstów wiedzy i informacji, badaniu powiązań w biznesie, polityce i nauce, czy wreszcie pracy na „surowych" tekstach publikowanych w Internecie w postaci informacji prasowych, artykułów, blogów, dokumentów itp. W odróżnieniu od komercyjnych wyszukiwarek internetowych, które najlepiej działają z małymi zestawami słów kluczowych, narzędzia CLARIN starają się w całości „zrozumieć" analizowane teksty i dlatego potrafią znaleźć w nich informacje związane z interesującym użytkownika tematem, nie wymagając wpisywania na chybił trafił różnych kombinacji wyrazów. Jest to szczególnie przydatne wtedy, kiedy np. w bardzo dużym zbiorze tekstów źródłowych

wyszukiwane są powiązania między wybranymi elementami (osobami, miejscami, instytucjami, przedsiębiorstwami itp.). Jak zapewnił prelegent, narzędzia CLARIN będą wsparciem w takich zadaniach związanych z przetwarzaniem języka, jak automatyczne streszczanie tekstów, wyszukiwanie w nich nazw własnych i słów kluczowych oraz analiza składniowa i morfologiczna. Tego rodzaju przetwarzanie pomoże badaczom np. w analizie dyskursu politycznego, społecznego czy reklamowego.

Po zakończeniu trzeciej sesji odbyła się krótka dyskusja na temat możliwości dalszego rozwoju projektu CLARIN.pl.

Konferencję podsumowała prof. Grabowska, która podziękowała prelegentom za ciekawe wystąpienia i podkreśliła ważkość podnoszonych tematów. Zwróciła także uwagę na konieczność współtworzenia międzynarodowych projektów typu DARIAH czy CLARIN w kontekście rozwoju humanistyki cyfrowej, wykorzystania cyfrowych technologii i narzędzi w badaniach i edukacji.

Ponieważ bogactwem big data, oprócz zgromadzonych informacji, jest ich wielowymiarowy kontekst, jego umiejętna analiza pozwala na pozyskiwanie olbrzymich zbiorów danych dla wszystkich branż od biznesu, przez kulturę, edukację i naukę, aż po świat mediów. Wyniki uzyskane z analizy big data tworzą wcześniej niedostępne źródła danych. Ich kreowanie może być postrzegane jako nowa faza rozwoju aplikacji IT (narzędzi i cyfrowych sieci wymiany informacji), współcześnie obejmując rozwój cloud computingu[4]. Te wzbogacające wiedzę informacje, przedstawione w trakcie spotkania, z pewnością okażą się przydatne w pracy naukowej i działalności zawodowej uczestników.

Julita Niedźwiecka-Ambroziak
uczestniczka studiów doktoranckich
z zakresu bibliologii i informatologii
prowadzonych na Wydziale Nauk Historycznych
Uniwersytetu Mikołaja Kopernika w Toruniu;
Biblioteka Wyższej Szkoły Bankowej w Toruniu

[4] W. Gogołek, *Informacyjny potencjał rafinacji zasobów sieciowych*. „Rocznik Bibliologiczno-Prasoznawczy" [online] 2014, tom 6/17 [dostęp 31 lipca 2016]. Dostępny w World Wide Web: http://www.ujk.edu.pl/ibib/studia/pdf/170/informacyjny_potencjal.pdf.