INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 1(31) • 2014

ISSN 1507-3858

Marta Tabakow, Jerzy Korczak, Bogdan Franczyk
Uniwersytet Ekonomiczny we Wrocławiu

BIG DATA – DEFINICJE, WYZWANIA I TECHNOLOGIE INFORMATYCZNE

Streszczenie: Big Data jako kompleks zagadnień informatycznych stanowi jedno z najważniejszych wyzwań współczesnego świata cyfrowego. W obecnych czasach, przy ciągłym napływie dużej ilości informacji pochodzących z różnych źródeł, a zatem o różnej charakterystyce, wymaga się wprowadzenia nowych technik analizy danych oraz rozwiązań technologicznych. W szczególności Big Data wymaga stosowania równoległego przetwarzania danych oraz odejścia od klasycznego schematu przechowywania danych. Zatem w niniejszej pracy dokonano przeglądu podstawowych zagadnień związanych z tematyką Big Data. Przedstawiono różne definicje Big Data, problemy badawcze i technologiczne oraz wyzwania dotyczące wolumenu danych, ich zróżnicowania, redukcji wymiaru, jakości danych i możliwości wnioskowania. Wskazano także dalszy kierunek prac w zakresie rozpoznania możliwości Big Data w różnych obszarach zarządzania.

Słowa kluczowe: Big Data, definicja Big Data, wyzwania Big Data, Hadoop, NoSql, Map Reduce, przetwarzanie równoległe.

DOI: 10.15611/ie.2014.1.12

1. Wstęp

Codziennie ma miejsce masowy napływ dużej ilości danych cyfrowych, pochodzących z różnych źródeł informacyjnych, takich jak: czujniki, dokumenty, fora internetowe. W ciągu ostatnich dwóch lat mamy do czynienia z takim przyrostem danych w postaci cyfrowej, że w tym czasie ich liczba osiągnęła 90% wszystkich zgromadzonych danych. Według IBM codziennie jest generowane 2,5 tryliona bajtów danych [www.ibm.com/software/data/bigdata], w tym przeważa zdecydowanie ilość danych semistrukturalnych (np. pliki typu XML wraz z odpowiednimi plikami XMLSchema), prawie strukturalnych (np. strumienie web clicks) ,niestrukturalnych (np. pliki tekstowe, PDFs, images videos). Obecnie olbrzymia ilość powstających informacji jest zjawiskiem globalnym, które dotyczy wszystkich podmiotów uczestniczących w różnych rynkach. Wzrost ilości danych cyfrowych na świecie spowodował, iż konwencjonalne techniki ich przetwarzania i przechowywania stały się nieadekwatne do

obecnych potrzeb. W związku z tym pojawił się trend zarządzania zasobami typu Big Data [Buhl i in. 2013, Zikopoulos i in. 2012], co stanowi wyzwanie nowoczesnego świata cyfrowego. Big Data jest przedmiotem dyskusji w badaniach naukowych oraz ma liczne zastosowania w praktyce. Termin Big Data jest trudny do zdefiniowania. W ogólnym znaczeniu dotyczy technologii gromadzenia i analizy danych o dużej objętości i złożoności. Dane pochodzą zarówno z tradycyjnych baz danych, np. funkcjonujących w przedsiębiorstwach, które zawierają tzw. dane wewnętrzne, jak i z innych źródeł (dokumentów, e-maili, blogów, mediów społecznościowych, różnego typu czujników elektronicznych, urządzeń lokalizacyjnych, np. GPS). Dane mają zarówno strukturę określoną, jak i nieokreśloną, co utrudnia ich dystrybucję i przetwarzanie za pomocą dostępnej infrastruktury informatycznej (architektury i narzędzi analitycznych) oraz metod obliczeniowych. Ta złożoność i różnorodność danych stwarza nowe wyzwanie dla analityków danych.

Wyszukując publikacje naukowe na przełomie listopada i grudnia 2013 r. według słowa kluczowego „big data" w bazie danych światowych wydawnictw naukowych Elsevier [www.elsevier.com] i Springer Journals [link.springer.com], wyświetla się 395 wyników, natomiast w bazie danych IEEE [http://ieeexplore.ieee.org] – 661 wyników. Wyniki te reprezentują wszystkie dziedziny, co wskazuje na duże zainteresowanie ze strony naukowców i dużą interdyscyplinarność tematu. W tym okresie obserwuje się również liczne konferencje naukowe i biznesowe na temat Big Data oraz zainteresowanie tym tematem komercyjnych czasopism IT, takich jak: „New York Times", „Forbes", „IBMSystems Magazine", które oferują specjalne wydania na ten temat. Duże koncerny, takie jak: IBM, HP, Teradata, Oracle, SAP, EMC, Amazon, Microsoft, Google, VMware, Cloudera, Hortonworks, Splunk, 10Gen, MapR, dostrzegają przyszłość w zbieraniu informacji i przetwarzaniu ich w taki sposób, aby ujawnić nową wiedzę o rynku, w którym funkcjonują, zobaczyć różne zależności i wzorce zachowań w społeczeństwie i środowisku [Korolov 2013]. Koncerny inwestują w gromadzenie danych, architekturę, platformy programistyczne, produkty i usługi, aplikacje analityczne itd.

Prawidłowa interpretacja danych odgrywa kluczową rolę w przedsiębiorstwach, a także znaczącą rolę w gospodarce krajowej, światowej i również społecznej. W przypadku zarządzania większa dostępność danych to dokładniejsza analiza i lepsze decyzje prowadzące do większej wydajności operacyjnej, obniżenia kosztów i zmniejszenia ryzyka [www.sas.com/big-data]. Dużą ilość informacji zawdzięcza się rozwojowi Internetu, coraz szerszej jego dostępności zarówno dla firm, jak i osób indywidualnych oraz rozwojowi nowoczesnych technologii i urządzeń do przekazu i przesyłu danych, takich jak telefony komórkowe, smartfony, tablety, GPS, czujniki mierzące i podające lokalizację, ruch, drgania, temperaturę, wilgotność, zawartość chemiczną w powietrzu (w szerokiej gamie urządzeń). To generacja innych, „nowych" danych – nowych w sensie dzisiejszego, łatwego dostępu do nich.

Celem niniejszej pracy jest przegląd literatury oraz przedstawienie propozycji definicji i atrybutów pojęcia Big Data. Opisane zostaną główne problemy badaw-

cze związane z gromadzeniem, przechowywaniem, wyszukiwaniem, udostępnianiem, przesyłaniem, analizą i wizualizacją tego typu danych, a także wyzwania związane z dostosowaniem technologii informatycznych do zarządzania zasobami Big Data, szczególnie w zakresie stosowania systemów rozproszonych do przetwarzania i przechowywania danych, takich jak narzędzia Hadoop i modele NoSql.

Analizę literatury przeprowadzono etapowo. Początkowo objęto badaniem publikacje naukowe z lat od 2009 do 2013, stanowiące 80% cytowanej literatury. Jednak podczas studiowania literatury dotyczącej Big Data należało sięgnąć do kilku publikacji z wcześniejszych lat: 1979-2009, gdzie również poruszano tematykę dużego przyrostu informacji i wyzwań z tym związanych. Autorzy do wyszukiwania literatury wybrali bazy danych: EBCSO Publishing, Springer, IEEE Xplore, Science Direct i Scopus. Zawężono przeszukiwania do publikacji pełnotekstowych w recenzowanych czasopismach naukowych, wydanych w latach 2009-2013. Ze względu na dużą interdyscyplinarność tematu wyszukiwano publikacje z różnych kategorii czasopism, takich jak: „business", „management", „computer science", „decision sciences", „engineering and technology", „social science". Wstępnie zastosowano słowo kluczowe „big data", następnie stosowano pary słów ze słowem „big data" i kolejnymi: „definition", „challenges", „opportunities", „technologies", „architecture", „new technologies", „management", „map reduce". Zastosowano procedurę iteracyjną przeszukiwania baz danych aż do wyodrębnienia oczekiwanych, ze względu na niniejszy przegląd, publikacji. W kolejnym etapie wyeliminowano opracowania spoza zakresu przeglądu. W tak powstałej bazie publikacji dokonano analizy abstraktów ze względu na istotne zagadnienia, a następnie pełnych tekstów publikacji naukowych. Dodatkowo przeanalizowano kilka najnowszych raportów kluczowych firm IT oraz opracowania międzynarodowych organizacji i firm konsultingowych.

Niniejsza praca została zorganizowana w następujący sposób. W kolejnym rozdziale zaprezentowana została definicja i charakterystyka najważniejszych atrybutów związanych z Big Data. Następnie przedstawiono składowe i wyzwania dotyczące architektury dedykowanej Big Data.

2. Big Data – ogólna charakterystyka

W związku z narastającym gwałtownie zainteresowaniem Big Data wielu naukowców i firm z branży IT podjęło próby rozpoznania i opisania tego terminu. Charakterystyka i definicje przez lata ewoluowały. Interpretując Big Data, należy się odnieść do nowych rozwiązań technologicznych dotyczących przetwarzania wielkich wolumenów danych o całkowicie innym charakterze (ilościowym i jakościowym) niż dotychczas [McKinsey 2011]. W artykule zaprezentowane zostały wybrane definicje oraz propozycja definicji wraz atrybutami, które charakteryzują Big Data. Przedstawiono wyzwania badawcze i technologiczne, m.in. istotność informacji, objętość danych, integracja danych, wnioskowanie.

2.1. Przegląd definicji

Poniżej przedstawiono kilka popularnych definicji Big Data. Jedna z pierwszych definicji Big Data została wprowadzona przez M. Cox i D. Ellsworth [Cox, Ellsworth 1997]. Autorzy traktują Big Data jako duże dane do analizowania, których liczbę należy maksymalizować w celu wydobycia wartości informacyjnych.

Inną propozycję definicji wysunął analityk pracujący dla Gartner w 2001 r., ówcześnie META Group (firmy analityczno-doradczej specjalizującej się w technologiach informacyjnych). Oparł ją na koncepcji trzech atrybutów w modelu „3V". Big Data charakteryzują atrybuty: objętość (***v**olume*), różnorodność (***v**ariety*), szybkość przetwarzania (***v**elocity*) [Doug 2001]. Następnie w roku 2012 firma Gartner [www.gartner.com] wprowadziła dodatkowe dwa wymiary odnoszące się do dużych danych: zmienność (*variability*) i złożoność (*complexity*).

Big Data to duża liczba danych, która wymaga zastosowania nowych technologii i architektur, tak by była możliwa ekstrakcja wartości płynącej z tych danych poprzez uchwycenie i analizę procesu, to sentencja, jaką przedstawiają autorzy publikacji [Katal i in. 2013].

Kolejną definicję Big Data prezentują Fan i Bifet [2012], opisując w niej Big Data jako termin oznaczający zbiory danych, którymi nie można zarządzać za pomocą obecnych metod eksploracji lub narzędzi programowych ze względu na duży rozmiar i złożoność danych.

IBM w 2013 r. definiuje Big Data jako różnorodne dane generowane z różnych źródeł, z dużą prędkością oraz w dużej ilości. IBM charakteryzuje Big Data za pomocą czterech atrybutów: objętość (***v**olume*), szybkość przetwarzania (***v**elocity*), różnorodność (***v**ariety*) oraz wiarygodność (*veracity*) [www.ibm.com].

SAS [www.sas.com] definiuje Big Data jako tendencje do poszukiwania i wykorzystania wartości biznesowej drzemiącej w dostępnych, coraz większych wolumenach danych, które charakteryzują się dużą zmiennością i złożonością. SAS opisując Big Data, zwraca uwagę na dodatkowe dwa atrybuty: zmienność (*variability*) oraz złożoność (*complexity*).

Na podstawie analizy definicji pojęcia Big Data można zauważyć, iż pojedynczo nie odzwierciedlają one w pełni tej problematyki. Nie można też wskazać na konkretną ilość bajtów, od których można mówić o dużych ilościach danych, gdyż tempo ich przyrostu jest zbyt wielkie [Cisco 2012]. W związku z tym poniżej przedstawiona została propozycja definicji Big Data.

Big Data to określenie stosowane dla takich zbiorów danych, które jednocześnie charakteryzują się dużą objętością, różnorodnością, strumieniowym napływem w czasie rzeczywistym, zmiennością, złożonością, jak również wymagają zastosowania innowacyjnych technologii, narzędzi i metod informatycznych w celu wydobycia z nich nowej i użytecznej wiedzy.

2.2. Cechy charakterystyczne Big Data

Termin Big Data jest charakteryzowany przede wszystkim kilkoma kluczowymi atrybutami, takimi jak: objętość (*volume*), różnorodność (*variety*), złożoność (*complexity*), strumień (*velocity*), zmienność (*variability*) i wartość (*value*). Big Data należy rozumieć jako techniki łączące rozwiązania z poniższych obszarów charakteryzujących dane. Poniżej przedstawiono definicje cech Big Data.

Objętość charakteryzuje się znaczącą dynamiką przyrostu danych, dla których wymagane są nowe technologie bazodanowe. Badania wskazują, że liczba danych do 2020 r. wzrośnie o 40% zeta bajtów, co oznacza 50-krotny wzrost od początku 2010 r.

Szybkość – dane napływające szybko, strumieniowo, które w związku z procesami biznesowymi wymagają dodatkowej mocy obliczeniowej do ich analizy w czasie rzeczywistym. Dane, które w związku z ograniczoną przepustowością sieci należy pobierać porcjami i wybierać tylko te, które mają istotną wartość informacyjną czy biznesową z punktu widzenia danej organizacji.

Różnorodność – dane pochodzą z wielu źródeł i często występują w różnych formatach i są zapisywane za pomocą różnych modeli oraz wyrażane w dowolnej formie, np.: liczbowo, tekstowo, obrazowo, dźwiękowo, oraz generowane w różny sposób.

Przykładem źródeł i różnego sposobu generowania danych są:

- dane tzw. wewnętrzne organizacji, przechowywane w bazach danych, np. transakcyjne, księgowe, kadrowe;
- dane ze źródeł zewnętrznych: internetowe, tworzone przez użytkowników Internetu; można tutaj wymienić: tweety, blogi, fora internetowe, sieci społecznościowe;
- dane z wszelkich transakcji [Chang i in. 2006]; ich źródłem mogą być sklepy, usługodawcy, instytucje finansowe;
- dane mające swoje źródło w placówkach służby zdrowia;
- dane z głębokiego Internetu (Deep Web data), które nie są indeksowane przez standardowe wyszukiwarki [He i in. 2007, Wei i in. 2010]; można wśród nich wyróżnić pochodzące z dynamicznych stron internetowych, odłączone treści na stronach internetowych, niezwiązane z innymi stronami, prywatne strony WWW, do których dostęp jest możliwy po zalogowaniu, lub strony o ograniczonym dostępie, gdzie aby przejrzeć ich zawartość, należy uprzednio wpisać sekwencję znaków;
- dane z wykresów, tworzone przez dużą liczbę węzłów informacyjnych i powiązań między nimi [Aggarwal, Wang 2010].

Zmienność – dane, których natężenie jest zmienne w czasie, a przepływy danych podlegają okresowym cyklom i trendom, a także szczytom, co związane jest również z dynamiką procesów i zmian gospodarczych czy politycznych. Przykładem może być sprzedaż produktów i usług w okresie Bożego Narodzenia, wzmożona aktywność w mediach społecznościowych towarzysząca wyborom parlamentarnym, nagłe „ruchy" na giełdzie czy okresowa rotacja portfeli inwestycyjnych.

Złożoność – złożoność danych jest ściśle związana z różnorodnością. Charakteryzuje się różnym uporządkowaniem danych. Są to m.in. dane o określonej strukturze, mające określony typ i format, dane o mieszanej strukturze, częściowo uporządkowane (*semi-struktured*, *„quazi" structured*), posiadające pewne właściwości organizacyjne, oraz dane niemające naturalnej struktury (*unstructured*), które należy zintegrować w celu odkrycia nieznanych relacji, powiązań i hierarchii. Do danych strukturalnych należą: numery telefonów, pesel, numer karty kredytowej – zawsze zawierają one zdefiniowaną liczbę pól. Dane o mieszanej strukturze to np. pliki XML, e-mail, Elektroniczna Wymiana Danych (Electronic Data Interchange, EDI). Natomiast dane niestrukturalne to: dokumenty, pliki wideo i zdjęcia. Ekstrakcja informacji z tych surowych treści oraz odpowiednio dobrane do nich metody są niezbędne do dalszego przetwarzania informacji przez algorytmy analizy [Chang i in. 2006; Labrinidis, Jagadish 2012].

Wartość – unikatowa wartość informacyjna ukryta w dużych i złożonych strukturach danych, dająca możliwość wyciągania nowych wniosków, które następnie przyczyniają się do wzrostu efektywności działania organizacji na różnych płaszczyznach. Można wskazać niektóre elementy działalności przedsiębiorstwa, na które ma wpływ wykorzystanie wartości, jakie niesie Big Data: efektywniejszy wewnętrzny model biznesowy, model doboru kadr, spersonalizowana oferta dla klientów, strategia marketingowa i konkurencyjna. Dodatkowo można zauważyć kolejną zaletę związaną ze stosowaniem Big Data w organizacji: dzięki Big Data można odpowiedzieć niemal natychmiast na stawiane pytania czy zdefiniowane problemy biznesowe. Wartość jest istotnym atrybutem Big Data. Może być rozumiana jako unikalna wiedza z naukowego punktu widzenia, jak i wartość informacyjna będąca korzyścią biznesową, mającą wpływ na obniżenie kosztów działalności organizacji czy na poprawę relacji biznesowych i zysków.

Inwestycje w Big Data, właściwie skierowane, mogą przyczynić się do osiągnięć naukowych i biznesowych na wielu płaszczyznach. Współczesne podmioty muszą sprostać tym wyzwaniom, aby mogły stanowić konkurencję na rynku zarówno krajowym, jak i międzynarodowym. Big Data to nowe wyzwanie i możliwości informacyjne [Jinchuan i in. 2013]. Prawidłowa interpretacja danych może odegrać kluczową rolę w gospodarce światowej i lokalnej, polityce społecznej oraz w przedsiębiorstwach. W zarządzaniu większa dostępność danych to dokładniejsza analiza i lepsze decyzje prowadzące do większej wydajności operacyjnej, obniżenia kosztów i zmniejszenia ryzyka [www.sas.com/big-data]. Z raportu *Big Data, Big Impact*, opublikowanego w 2012 r. przez The World Economic Forum, wynika że dane stanowią nową klasę ekonomicznych aktywów. Jednak, aby w pełni wykorzystać możliwości Big Data, należy traktować zagadnienie kompleksowo i opracowywać koncepcję architektoniczną, czyli szkielet działania systemu informatycznego, od podstaw dla konkretnej branży czy zastosowania.

Analizując obszary wykorzystania możliwości Big Data, trudno wymienić wszystkie. Niewątpliwie są to: usługi finansowe, edukacja, zdrowie, rolnictwo, bezpieczeństwo, „inteligentne" zarządzanie, planowanie miejskie, logistyka w transporcie, modelowanie środowiska, oszczędzanie energii, wody. Niektóre przykłady zastosowań można znaleźć w [Labrinidis, Jagadish 2012].

Definicja Big Data w dalszym ciągu ewoluuje w wyniku powstawania nowych źródeł informacji, rozszerzania się zakresu zastosowań i rozwoju technologii informacyjno-komunikacyjnych.

2.3. Systematyka Big Data

Można wstępnie dokonać próby syntezy prac naukowych i raportów branży IT w obszarze tematycznym Big Data. Główną grupę stanowią publikacje ogólne, przedstawiające charakterystykę i definicję Big Data, często przekrojowo wskazujące wyzwania i możliwości, jakie się w związku z tym pojawiają, zarówno dla biznesu, jak i dla nauki, w różnych branżach. Kolejne publikacje związane są z technologiami przetwarzania i przechowywania danych [Boja i in. 2012]. Można w tym obszarze wyróżnić taką technologię, jak Cloud Computing stanowiącą możliwość korzystania z zasobów informatycznych poza siedzibą przedsiębiorstwa. Następnie zestaw narzędzi Hadoop, stanowiący uzupełnienie istniejącej infrastruktury informatycznej o możliwość analizy w czasie rzeczywistym dużych zbiorów danych [Zikopoulos i in. 2012] oraz platforma Map Reduce, związana z rozproszonym systemem plików [Sakr i in. 2013]. Kolejna grupa publikacji dotyczy badań związanych ze wstępnym przetwarzaniem danych i sposobem ich analizowania, śledzenia, poszukiwania wzorców w czasie rzeczywistym oraz przetwarzania zdarzeń, łączących dane z różnych źródeł [Sakr i in. 2013]. Można tu wyróżnić metodę złożonego przetwarzania zdarzeń (Complex Event Procesing, CEP) [Mozafari i in. 2013] oraz prace opisujące narzędzia i metody do przetwarzania danych. Szczególną grupą publikacji są opracowania dotyczące przetwarzania danych pozyskanych z różnych źródeł, występujących w różnych formatach [Bostock i in. 2013]. Szczególną uwagę zwraca się na metody Web Miningowe [Chang i in. 2006], służące do odkrywania wzorców w sieci, i metody analizy dokumentów o złożonej strukturze [Bostock i in. 2013].

2.4. Wyzwania technologiczne dla Big Data

Wyzwania badawcze wynikają z charakterystyki Big Data. Związane są z ilością danych, z ich złożonością, różnorodnością oraz liczbą źródeł informacji [Jinchuan i in. 2013]. Wyzwania badawcze można podzielić na technologiczne i związane z samymi danymi. Wyzwania technologiczne to przede wszystkim opracowanie innowacyjnej architektury, przy czym architekturę rozumie się jako szkielet dla całego procesu związanego z wykorzystywaniem danych, począwszy od wyznaczenia źródeł danych, które są interesujące z punktu widzenia biznesowego, poprzez ich

pobieranie, gromadzenie, wstępne przetwarzanie, rozdzielanie, analizę, modelowanie do wnioskowania.

Obecnie Big Data to „nowe" dane z różnych źródeł. Potrzebna jest zatem identyfikacja istotnych źródeł danych. Nasuwają się więc pytania: Z jakich źródeł czerpać dane? Jakie źródła będą istotne dla realizacji celów biznesowych? Jak identyfikować te źródła? Czy można je identyfikować automatycznie i czy te narzędzia będą personalizowane? Narzędzia takie powinny być częścią architektury przeznaczonej do zarządzania Big Data. Identyfikacja istotnych źródeł danych prowokuje kolejne pytania i wyzwania, które zostaną poruszone w dalszej części rozdziału.

Identyfikacja istotnych informacji to ważny aspekt zarządzania danymi Big Data. Istnieje konieczność określenia filtrów danych, by z jednej strony nie pobierać nadmiernych, niepotrzebnych danych i nie generować zbędnych kosztów, a z drugiej strony, by nie pominąć tych istotnych. Ponadto wielkim wyzwaniem jest automatyczne generowanie metadanych, opisujących dane.

Duże ilości danych napływające z nowoczesnych urządzeń, czujników, aplikacji internetowych, z powodu swojego wolumenu, przerosły możliwości dzisiejszych baz danych (Data Base Management Systems, DBMS). Również trudne stało się przetwarzanie zestawów danych w sposób tradycyjny i prezentacja wyników np. za pomocą wykresu. Wobec powyższego wydłużył się czas oczekiwania na wyniki, natomiast systemy transakcyjne wymagają krótkiego czasu reakcji, niezależnie od wielkości zbioru danych. W rzeczywistości oczekiwane jest przetwarzanie informacji na bieżąco w czasie rzeczywistym.

Strumieniowy napływ danych to wyzwanie związane z koniecznością obsługi „nowych" danych i ich aktualizacji. Prędkość i ilość danych w obecnych strumieniach danych wymaga od nowoczesnych systemów niezwłocznej (bieżącej) ich obsługi. Szybkość napływu danych to przede wszystkim wyzwania dla platformy zarządzania stosem danych. Związane to jest zarówno z warstwą przechowywania, jak i przetwarzania zapytań. Obie muszą być bardzo szybkie i skalowalne. Technologie te są przez ostatnie lata nieustająco rozwijane, jednakże w dalszym ciągu ograniczone.

Obecnie dane zwykle pochodzą z różnych źródeł. Różnorodność jest charakterystyczną cechą dużych danych ze względu na szeroki zakres źródeł. W związku z tym dane pojawiają się w różnych formatach i modelach. Zachodzi potrzeba wdrożeń nowych technologii i nowych narzędzi do integracji tych danych, tak by znaleźć powiązania i zależności między pozornie różnymi obszarami, w których znajdują się dane. Przykładem może być np. powiązanie pogody z decyzją inwestorów na giełdzie. W ramach różnorodności można mówić również o strukturze danych, a raczej o nieuporządkowaniu danych, co stanowi odrębne wyzwanie. Istnieją technologie do radzenia sobie z różnymi typami danych, jednak w dalszym ciągu ich bezproblemowa integracja pozostaje wyzwaniem.

2.5. Wyzwania związane z danymi

Razem z ogólnymi wyzwaniami dotyczącymi Big Data identyfikowane są wyzwania związane z danymi, ich źródłami, ilością, wielowymiarowością, jakością, informacją, jaką można z nich uzyskać, oraz wartością biznesową dającą się przełożyć na konkretny cel organizacji.

Dane klasy Big Data pochodzą z różnych źródeł, zawierają różnego rodzaju dane i występują w różnych formatach. W części dane te są nieistotne z punktu widzenia celu biznesowego organizacji oraz zawierają błędne lub nieprawdziwe informacje. Poważnym wyzwaniem jest zatem ocena ich jakości. Przykładem są treści z sieci społecznościowych, gdzie nie wiadomo, czy użytkownicy wpisują w nich prawdę i czy nie tworzą w nich tzw. drugiego życia, kreując na portalach nowy, inny obraz swojej osoby. Banki i inni przedsiębiorcy, opierając się na fałszywych danych, są w związku z tym narażeni na podejmowanie błędnych decyzji, co w konsekwencji naraża ich na dodatkowe koszty lub straty, związane np. ze źle przygotowaną kampanią marketingową czy udzieleniem kredytu osobie niezdolnej do jego spłaty.

Szczególnym problemem zarządzania danymi Big Data jest redukcja wymiarowości danych w kontekście ich analizy i wizualizacji – wizualna redukcja danych (Visual Data Reduction) [Keim i in. 2010]. Jest to jedno z bardziej wymagających wyzwań analizy danych, gdyż rekordy danych są wielowymiarowe, często wielomiliardowe. Wizualizacja tak dużych zestawów danych jest często niemożliwa i z reguły nieczytelna dla użytkownika. Ponadto systemy komputerowe oraz technologie ograniczają zdolność do szybkiego działania przy tworzeniu wizualizacji. W związku z tym koncerny i naukowcy na świecie pracują nad udoskonaleniem zarówno algorytmów redukcji wymiarowości danych, jak i innowacyjnymi podejściami do hierarchicznego zarządzania danymi i skalowalnością ich struktur w celu ich wizualizacji w 2D lub 3D, tak by ich graficzna interpretacja była możliwa. Dodatkowo, oprócz technicznych i obliczeniowych problemów, wskazać można inne aspekty badawcze, takie jak: Kiedy należy wykorzystywać techniki wizualizacji? Która metoda wizualizacji jest odpowiednia dla danego zbioru danych? W jakim zakresie i jak dokonywać oceny jakości zredukowanego zbioru danych?

Aby w pełni skorzystać z potencjału Big Data, organizacja staje przed wyzwaniem, jakim jest ponowna weryfikacja celów biznesowych i ustalenie celów analizy. W zależności od rodzaju prowadzonej działalności cele mogą się różnić. Mając określone cele biznesowe, wyznacza się zakres źródeł oraz rodzaje danych, które są niezbędne w procesie analitycznym. Celami biznesowymi mogą być m.in.: obniżenie kosztów działalności, wzrost zysków, nowi konsumenci, nowe rynki zbytu, nawet wizerunek.

Jak zrozumieć dużą ilość danych? Jak przełożyć dane na informacje? Które są ważne? Jak wyciągać wnioski? Poważnym problemem w dobie Big Data staje się odpowiedź na powyższe pytania. Do zrozumienia danych oraz wyciągnięcia z nich od-

powiednich wniosków potrzebne są zarówno nowe narzędzia do analizy danych, jak i wysokie umiejętności analityczne osób zajmujących się przetwarzaniem informacji. Zrozumienie danych stało się przedmiotem badań i inwestycji wśród organizacji. Ponadto reakcji na napływ informacji oczekuje się w czasie rzeczywistym. Pojawiły się całe gałęzie przemysłu zajmujące się gromadzeniem danych i ich zrozumieniem [www.ibmsystemsmag.com/power/infrastructure/Linux/affordable_analytics]. Kadra zarządzająca w sektorze MŚP widzi konieczność do zbadania potencjału Big Data.

Badania dowodzą, że narzędzia powinny zostać zmodyfikowane, a metody opracowane lub „odświeżone" jako przeznaczone „nowemu" typowi danych [Leishi i in. 2012; Keim i in. 2010; Thomas, Cook 2006], aby w pełni skorzystać z potencjalnej wartości wynikającej z Big Data.

3. Architektury informatyczne dedykowane Big Data

Jak już wspomniano, charakterystyka Big Data wiąże się z wielkością zbiorów danych, szybkością napływu nowych i dużą ich różnorodnością. W związku z tym dane są trudne do analizowania i wnioskowania. Pojawiają się więc nowe wyzwania dla firm związane z infrastrukturą IT. Zazwyczaj dla organizacji wiąże się to z dodatkowymi nakładami inwestycyjnymi na zakup dodatkowego sprzętu, np. serwerów, pamięci masowej oraz oprogramowania. Proponowanymi odpowiedziami na wyzwania są m.in. możliwości zestawu oprogramowania Apache Hadoop [http://hadoop.apache.org], projekt Stratoshpere [http://stratosphere.eu/] oraz model Cloud Computing [Jadeja, Modi 2012].

3.1. Cloud Computing

Cloud Computing to pojęcie używane do określenia skalowalnej platformy zawierającej sprzęt IT wraz z oprogramowaniem, która dostępna jest u zewnętrznego operatora jako usługa dostarczana za pośrednictwem Internetu. Cloud Computing oznacza również system rozproszenia, zdolność uruchamiania programu lub aplikacji na wielu połączonych komputerach w tym samym czasie lub dynamiczną obsługę danego żądania, polegającą na przydzieleniu zadania do jednego z dostępnych serwerów. Cloud Computing jest wsparciem dla rozwoju firm oraz wygodnym rozwiązaniem. Umożliwia działanie zasobów informatycznych firmy w tzw. chmurze, gdzie użytkownik, a w tym przypadku firma, nie musi posiadać w swoich zasobach urządzeń (np. serwerów) ani oprogramowania, natomiast wszystko podnajmuje od innych firm. Ma to szczególne zastosowanie dla firm z sektora MŚP, gdzie ma duże znaczenie ograniczenie kosztów. Istnieje zbieżność Cloud Computing i Big Data, ponieważ model Cloud Computing zapewnia nieograniczone zasoby na żądanie, co odpowiada na wyzwanie wzrastającej objętości danych. Technologia Cloud Computing umożliwia rozwój firmom, które chcą wprowadzić rozwią-

zania biznesowe w oparciu o Big Data oraz zbierać i analizować dane niestrukturalne (zob. punkt 2.1), stanowiące obecnie 80% wszystkich danych na świecie. Cloud Computing jest rozwiązaniem do przetwarzania, przechowywania, dystrybucji i przenoszenia danych Big Data z „chmury do chmury", gdzie pod pojęciem „chmury" można rozumieć rozproszoną przestrzeń dyskową i zestaw serwerów (ciągle powiększających się) oferujących usługi informatyczne dowolnego typu.

Analiza możliwości związanych z zastosowaniem Cloud Computing, w przypadku zarządzania danymi Big Data, pozwala zauważyć jego znaczącą rolę zarówno w dużych, jak i małych przedsiębiorstwach oraz w różnych branżach. Cloud Computing stwarza możliwości przechowywania i analiz dużej ilości danych, również doraźnie, na żądanie. W związku z rosnącym zapotrzebowaniem na rozwiązania Cloud Computing dostawcy dostarczający tego typu usługi szukają rozwiązań dla ulepszenia architektury działania „chmury", zwiększenia pojemności pamięci, wydajności, szybkości przetwarzania informacji oraz elastyczności usług informatycznych.

3.2. Architektury dla zasobów

W ramach umożliwienia analizy dużej ilości danych oraz agregacji uzyskanych wyników, współczesne architektury przeznaczone Big Data tworzone są jako architektury wielowątkowe, równoległego przetwarzania [Shvachko i in. 2010, Changqing i in. 2012]. Ogólna koncepcja obowiązująca w zaproponowanych modelach przetwarzania informacji związana jest z klasycznym algorytmicznym podejściem typu „dziel i zwyciężaj", w których wejściowy problem (o wysokiej złożoności) dzielony jest na podproblemy o wyraźnie zredukowanej złożoności.

Pod hasłem „architektura" rozumie się połączenie sprzętu i odpowiedniego oprogramowania, które wspierają (umożliwiają) przetwarzanie danych Big Data. Oczywiście, należy zauważyć szereg kwestii, które należy rozwiązać w ramach tego typu koncepcji:

- sposób partycjonowania danych – wymagany jest podział informacji nie tylko pod względem objętości (redukcja ilości danych), ale również typu informacji;
- sposób przechowywania danych w celu ich rozbicia (partycjonowania), np. realizacja algorytmów indeksacji danych w bazach danych w celu przyspieszenia wyszukiwania informacji. W ramach problemu Big Data wprowadzono nową koncepcję indeksowania rekordów – indeksowanie fraktalne (*fraktal tree indexing*). Przykładowym rozwiązaniem bazodanowym, wykorzystującym indeksowanie fraktalne, jest TokuDB [www.tokutek.com];
- optymalizacja użycia jednostek obliczeniowych w celu uniknięcia „wąskich gardeł". Prowadzi się wiele prac badawczych, które dotyczą tworzenia optymalnych modeli przepływu informacji w sieci;
- agregacja informacji – w zależności od złożoności docelowego problemu może stać się dużym wyzwaniem badawczym.

Istnieje wiele różnych rozwiązań architektonicznych proponujących rozwiązania powyżej wymienionych zagadnień. Jedną z najczęściej cytowanych (wykorzystywanych) „open-source-owych" architektur tego typu jest projekt Apache Hadoop [hadoop.apache.org/index.html].

3.3. Apache Hadoop

Apache Hadoop to zestaw oprogramowania typu *open-source*, które umożliwia przetwarzane rozproszone dużych zestawów danych w klastrach serwerów [Venner 2009]. Jest wysoce skalowalne, umożliwia skalowanie z jednego do tysiąca komputerów, z bardzo wysokim stopniem odporności na uszkodzenia. Na architekturę Apache Hadoop składają się następujące moduły:

- Hadoop Common – zawiera biblioteki i narzędzia niezbędne do pracy modułów Hadoop;
- Hadoop Distributed File System (HDFS) – system zarządzania plikami w środowisku rozproszonym, który przechowuje i zarządza plikami na jednostkach rozproszonych i zapewnia wysoką, łączną przepustowość pomiędzy nimi;
- Hadoop YARN – platforma zarządzania zasobami odpowiedzialna za zarządzanie zasobami i obliczeniami w „klastrach" (podzbiory jednostek obliczeniowych) i wykorzystywania ich do harmonogramowania zadań użytkowników;
- Hadoop MapReduce – model programistyczny, przeznaczony przetwarzaniu dużej ilości danych. Moduły Hadoop MapReduce i HDFS [Changqing i in. 2012] wzorowane są na rozwiązaniach Google: Google's MapReduce i Google File System (GFS). Z punktu widzenia projektanta systemów kluczowy jest moduł MapReduce, ponieważ za pomocą odpowiedniego interfejsu programistycznego można zaprojektować procesy mapowania i redukcji. MapReduce jest systemem przeznaczonym dla tworzenia aplikacji działających jednocześnie na tysiącach komputerów. Główną jego zaletą jest umożliwienie łatwego rozproszenia operacji. Zakładając, że każda z operacji „map" jest niezależna od pozostałych, może być ona realizowana na osobnym serwerze.

Platforma Hadoop umożliwia uruchamianie aplikacji w systemach z tysiącami węzłów z udziałem tysięcy terabajtów danych. Rozproszony system plików umożliwia szybki transfer danych pomiędzy węzłami i zapewnia ciągłość w działania w przypadku awarii. Główne cechy platformy Hadoop dla rozwiązań związanych z wyzwaniami Big Data to:

- skalowalność, umożliwiająca dodawanie nowych węzłów w razie potrzeby, bez konieczności zmiany formatów danych, jakie są ładowane;
- opłacalność, ponieważ umożliwia równoległe obliczenia na serwerach, co w efekcie obniża koszty za każdy terabajt pamięci;
- elastyczność, ponieważ daje możliwość pobierania wszystkich rodzajów danych, z dowolnych źródeł, bez względu na ich strukturę. Zapewniająca ponadto

łączenie danych z wielu źródeł w dowolny sposób, umożliwiając przy tym wykonywanie dokładnych analiz;
– odporność na uszkodzenia. Oprogramowanie działa nawet wtedy, gdy jeden z węzłów został uszkodzony, przekierowując zadanie do innej jednostki.

Z punktu widzenia analityka danych kluczowa w architekturach dedykowanych Big Data jest możliwość własnej implementacji, na bazie celowej – w kontekście danego problemu analitycznego – implementacji modułów typu MapReduce, gdzie ulokowana jest warstwa zaawansowanej analizy danych, lub zastosowania tzw. technik sztucznej inteligencji.

3.4. NoSQL

Zagadnienia związane ze złożonością danych, szczególnie danych niestrukturalnych lub półstrukturalnych, powodują wyzwania w zakresie ich efektywnego przechowywania oraz przeszukiwania w bazie danych. Schemat relacyjnej bazy danych, sztywno zdefiniowanej, nie odpowiada danym typu Big Data. Zaprojektowany został więc mechanizm przechowywania danych NoSQL. Jest to „magazyn" danych zapewniający przechowywanie i pobieranie danych w nieograniczony sposób. Różni się on od powszechnie stosowanego modelu relacyjnego. Mimo iż to rozwiązanie jest powszechnie dostępne, napotyka się bariery w jego wykorzystywaniu w przedsiębiorstwach, szczególnie tych z sektora MŚP. Bariery te związane są ze skomplikowanym użyciem NoSQL. Mianowicie wykorzystywane są niskopoziomowe języki zapytań, co wiąże się z zatrudnieniem wykwalifikowanej obsługi. Dodatkowo brak jest intuicyjnych czy standardowych interfejsów wspomagających użytkowników.

Dzięki wdrożeniu NoSQL mogą zostać osiągnięte takie korzyści, jak poprawa zrozumienia danych, możliwość gromadzenia danych niestrukturalnych, skalowalność i elastyczność bazy oraz potencjalna możliwość tworzenia unikalnych modeli biznesowych opartych na wielu danych, których wcześniej nie można było uwzględnić w analizach. Przykładem bazy danych opartej na modelu NoSQL jest Cassandra.

3.5. Moduł ETL i ELT

Wybór odpowiedniego procesu ETL czy ELT zależy w głównej mierze od konfiguracji infrastruktury oraz od celów jakim ma służyć dalsze przetwarzanie i analiza danych. Model ETL (Extract, Transform and Load) kolejno oznacza pozyskanie danych, następnie różne transformacje danych mające na celu ich przekształcenie, by w łatwy sposób można je było przyporządkować do tabeli (np. zmianę formatu w którym są zapisane) w hurtowni danych oraz załadowanie danych. Przykładami struktur zawierających narzędzia ETL są systemy Business Intelligence. Technologia ETL była powszechnie stosowana w czasach, gdy przechowywanie danych było kosztowne, a przetwarzania czasochłonne, kiedy raporty i sprawozdania były

ściśle uwarunkowane stosowaniem modelu ETL.Obecnie wzrosła zdolność sprzętowa do przechowywania danych oraz wydajniejsze są również systemy do zarządzania bazami danych. Model ELT (Extract, Load, Transform and Load) zmienia kolejność procesu, ładuje surowe dane do magazynu danych i przekształca je wewnątrz. ELT staje się problemem, jeśli hurtownia danych jest zbytnio obciążona. Model ELT wymaga zmiany podejścia wobec tradycyjnych metod projektowania. Pozwala on na prześledzenie historii danych aż do źródła, dzięki przechowywaniu danych surowych, co w dalszej mierze pozwala na zaobserwowanie unikalnych zależności i powiązań, co dotychczas nie było możliwe.

Dzięki dostępności narzędzi, które mogą wspomóc implementację ELT, przedsiębiorstwa mogą osiągnąć korzyści ze stosowania tego modelu, szczególnie w zakresie poszukiwania unikalnych zależności pomiędzy danymi Big Data i wyciągania unikalnych wniosków w wyniku analizy tych danych.

4. Zakończenie

Big Data stanowi jedno z najważniejszych wyzwań współczesnego świata cyfrowego. Możliwości przetwarzania dużych ilości danych o różnym typie i dużej złożoności, pochodzących z różnych źródeł informacyjnych, znajdują zastosowanie w wielu dziedzinach: typowo naukowo-badawczych i komercyjnych. Zastosowania komercyjne dotyczą praktycznie każdej branży, jako że w sposób pośredni lub bezpośredni polityka firm uzależniona jest od dostępu do informacji i analizy odpowiednich danych. W niniejszej pracy został zaprezentowany przegląd prac dotyczących problematyki i charakterystyki Big Data. Przedstawiono kilka głównych definicji Big Data oraz zaproponowano własną definicję. Obszernie omówiono wyzwania związane z Big Data. Ponadto scharakteryzowano technologie informatyczne wykorzystywane w zarządzaniu zasobami Big Data, takie jak przechowywanie danych w „chmurze", systemy rozproszone w przetwarzaniu informacji.

Na podstawie przeglądu literatury można wywnioskować, że technologia ta stwarza nowe możliwości dla badaczy. Duża ilość różnorodnych danych z wielu źródeł to czynnik, który ma decydujący wpływ na jakość wyników w badaniach naukowych. Lepsze, w sensie jakościowym, bazy wiedzy determinują wyciąganie unikalnych, ze względów naukowych, wniosków. Ponadto nowe technologie gromadzenia i przetwarzania danych umożliwią prowadzenie interdyscyplinarnych badań naukowych. Dla przedsiębiorców natomiast Big Data to również nowe perspektywy rozwoju firmy. Do najważniejszych możliwości w tym zakresie zaliczyć można: uzyskanie przewagi konkurencyjnej na rynku poprzez tworzenie dopasowanych modeli predykcyjnych, w wyniku których klientowi proponuje się produkty i usługi idealnie dopasowane oraz optymalizację zarządzania przedsiębiorstwem, co prowadzi do zwiększenia zysków. Przyszłość informatyzacji w przedsiębiorstwach oraz utrzymanie korzystnej pozycji rynkowej oparte będą na technikach wykorzystujących Big Data.

W kolejnej pracy związanej z zarządzaniem zasobami typu Big Data planowany jest przegląd praktycznych możliwości wykorzystania dużych danych z punku widzenia realizacji konkretnych celów biznesowych.

Literatura

Aggarwal C.C., Wang H., 2010, *Graph Data Management and Mining: A Survey of Algorithms and Applications*, „Managing and Mining Graph Data", Series: Advances in Database Systems, Vol. 40, Springer, s. 13-68.

Bandler J., Grinder J., 1979, *Frogs into Princes: Neuro Linguistic Programming*, „Real People Press".

Buhl H., Röglinger M., Moser F., Heidemann J., 2013, *Big Data – Ein (ir-)relevanter Modebegriff für Wissenschaft und Praxis?*, „Wirtschaftsinformatik & Management", Springer, s. 24-31.

Boja C., Pocovnicu A., Batagan L., 2012, *Distributed Parallel Architecture for „Big Data"*, „Informatica Economica", Bucharest, Romania, vol. 16, issue 2, s. 116-127.

Bostock M., Ogievetsky V., Heer J., 2011, *D3: Data-driven documents*, „IEEE Transaction on Visualization & Computer Graphics", IEEE, vol. 17, issue 12, s. 2301-2309.

Cisco, *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011-2016*, 2012, http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html, 30.11.2013.

Chang C., Kayed M., Girgis M.R., Shaalan K.F., 2006, *A survey of web information extraction systems*. „IEEE Transactions on Knowledge and Data Engineering", IEEE, vol. 18, issue 10, s. 1411-1428.

Changqing J., Yu L., Wenming Q., Awada U., Keqiu L., 2012, *Big Data Processing in Cloud Computing Environments*, 12th International Symposium on Pervasive Systems, Algorithms and Networks (ISPAN), San Marcos, IEEE, s. 17-23.

Cox M., Ellsworth D., *Managing Big Data for Scientific Visualization*, 1997, ACM SIGGRAPH '97 Course #4, Exploring Gigabyte Datasets in Real-Time: Algorithms, Data Management, and Time-Critical Design, Los Angeles.

Doug L., 2001, Data *Management: Controlling Data Volume, Velocity, and Variety*, „Application Delivery Strategies", META Group (currently with Gartner).

Zikopoulos P., Eaton C., deRoos D., Deutsch T., Lapis G., 2012, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw Hill, USA.

Fan W., Bifet A., 2012, *Mining big data: current status, and forecast to the future*, „ACM SIGKDD Explorations Newsletter", SIGKDD Explorations, ACM, New York, USA, vol. 14, issue 2, s. 1-5.

He B., Patel M., Zhang Z., Chang K.C.C., 2007, *Accessing the deep web*, „Communications of the ACM", ACM, New York, USA, vol. 50, issue 5, s. 94-101.

Jadeja Y., Modi K., 2012, *Cloud computing - concepts, architecture and challenges*, 2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET), IEEE, Kumaracoil, India.

Jinchuan C., Yueguo C., Xiaoyong D., Cuiping L., Jiaheng L., Suyun Z., Xuan Z., 2013, *Big data challenge: a data management perspective*, „Frontiers of Computer Science", SP Higher Education Press, vol. 7, issue 2, s. 157-164.

Katal A., Wazid M., Goudar R.H., 2013, *Big Data: Issues, Challenges, Tools and Good Practices*, 2013 Sixth International Conference on Contemporary Computing (IC3), IEEE, Noida, s. 404-409.

Keim D., Kohlhammer J., Ellis G., Mansmann F., 2010, *Mastering The Information Age – Solving Problems with Visual Analytics*, Eurographics Association, Germany.

Korolov M., 2013, 15 most powerful Big Data companies, Network World.

Labrinidis A., Jagadish H., 2012, *Challenges and opportunities with big data*, Proceedings of the VLDB Endowment, VLDB Endowment, vol. 5, issue 12, s. 2032-2033.
Leishi Z., Stoffel A., Behrisch M., Mittelstadt S., Schreck T., Pompl R., Weber S., Last H., Keim D., 2012, *Visual analytics for the big data era — A comparative review of state-of-the-art commercial systems*, 2012 IEEE Conference on Visual Analytics Science and Technology, IEEE, Seattle, WA, s. 173-182.
McKinsey Global Institute, 2011, *Big data: The next frontier for innovation, competition, and productivity*, http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation, 30.11.2013.
Mozafari B., Zeng K., D'antoni L., Zaniolo C., 2013, *High-Performance Complex Event Processing over Hierarchical Data*, ACM Transactions on Database Systems (TODS), ACM, New York, USA, vol. 38, issue 4.
Rhodes R., 2013, *Finding Big Data's Sweet Spot*, IBM Systems Magazine, http://www.ibmsystemsmag.com/, 2.12.2013.
Sakr S., Liu A. Fayoumi A.G., 2013, *The Family of MapReduce and Large-Scale Data Processing* System, ACM Computing Surveys (CSUR), ACM, New York, USA, vol. 46, issue 1.
Shvachko K., Kuang H., Radia S., Chansler R., 2010, *The Hadoop Distributed File System*, 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), IEEE, Incline Village, NV, s. 1-10.
Thomas J.J., Cook A.K., 2006, *Visual Analytics Agenda*, „IEEE Computer Graphics and Applications", IEEE, s. 10-13.
Venner J., 2009, *Pro Hadoop*, Apress.
Wei L., Xiaofeng M., Weiyi M., 2010, *ViDE: A Vision-Based Approach for Deep Web Data Extraction*, IEEE Transactions on Knowledge and Data Engineering, IEEE, s. 447-460.
Wong P.C., Thomas J., 2004, *Visual analytics*, IEEE Computer Graphics and Applications, IEEE, s. 20-21.
World Economic Forum, *Big Data, Big Impact: New Possibilities for International Development*, Geneva 2012, http://www.weforum.org/, 3.12.2013.
Zikopoulos P., deRoos D., Parasuraman K., Deutsch T., Corrigan D., Giles J., 2013, *Harness the Power of Big Data: The IBM Big Data Platform*, McGraw-Hill, USA.

BIG DATA − DEFINITIONS, CHALLENGES AND INFORMATION TECHNOLOGIES

Summary: Big Data as a complex IT issues, is one of the most important challenges of the modern digital world. At the present time, the continuous inflow of a large amount of information from different sources, and thus with different characteristics, requires the introduction of new data analysis techniques and technology. In particular, Big Data requires the use of parallel processing and the departure from the classical scheme of data storage. Thus, in this paper we review the basic issues related to the theme of Big Data: different definitions of „Big Data" research and technological problems and challenges in terms of data volume, their diversity, the reduction of the dimension of data quality and inference capabilities. We also consider the future direction of work in the field of exploration of the possibilities of Big Data in various areas of management.

Keywords: Big Data, Big Data definition, challenges of Big Data, Hadoop, NoSql, Map Reduce, parallel processing.