# *Big Data* w inżynierii

## Big Data in engineering

Materiały z XX SKWPWiE, Jurata 2016 r.
DOI: 10.17814/mechanik.2016.7.183

**RAFAŁ RACZKO** *

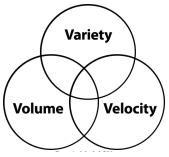
**W pracy omówiono zagadnienia związane z możliwością i perspektywami wykorzystania technologii *Big Data* w inżynierii. Zdefiniowano pojęcie *Big Data*. Omówiono wybraną metodę przetwarzania danych w technologii *Big Data*. Przedstawiono możliwości wykorzystania *Big Data* w inżynierii.**
**SŁOWA KLUCZOWE: *Big Data*, hadoop, analiza danych**

*In the following paper issues related to the possibility and perspectives of using Big Data technology in engineering were presented. The concept of Big Data was defined. The chosen method of processing data in Big Data technologies was discusses. The possibility of using Big Data engineering was shown.*
*KEYWORDS: Big Data, hadoop, data analysis*

Jeszcze kilkanaście lat temu nikt nie słyszał o *Big Data*. Aktualnie technologia ta rozwija się w sposób wyjątkowo energiczny i jest wykorzystywana praktycznie na całym świecie. Stanowi jedno z najważniejszych wyzwań współczesnego świata cyfrowego. *Big Data* umożliwia szybką analizę i wnioskowanie na bazie dużych ilości danych, m.in. w procesach projektowania, wytwarzania czy eksploatacji urządzeń i maszyn. Po raz pierwszy termin *Big Data* został użyty pod koniec lat dziewięćdziesiątych ubiegłego stulecia. Od tego czasu technologia ta regularnie zyskuje na popularności. Jednak pomimo upływu czasu wciąż nie istnieje jedna, powszechnie stosowana definicja *Big Data*. Dopiero w ciągu ostatnich kilku lat powstały istotne pozycje literaturowe oraz artykuły naukowe w tej dziedzinie. Przegląd literatury wskazuje, że temat jest wysoce interdyscyplinarny.

### Co to jest *Big Data*?

Pojęciem *Big Data* często określa się niesłusznie każdy duży zbiór danych. Wydaje się zatem, że dalej pojęcie to nie jest w pełni rozumiane. W rzeczywistości z *Big Data* mamy do czynienia dopiero wówczas, gdy tradycyjna eksploracja danych (*data mining*) nie jest w stanie wyszukać ukrytych wzorców i znaczeń [1]. McKinsey Global Institute definiuje *Big Data* następująco: „*Big Data* odnosi się do zbiorów danych, których rozmiar uniemożliwia przechwytywanie, przechowywanie, zarządzanie i analizowanie przez typowe narzędzia baz danych" [2]. Inna charakterystyka *Big Data* zaproponowana przez Douga Laneya z firmy Gartner w roku 2001, została pokazana na rys. 1.




Rys. 1. Model 3V

Objętość (*volume*) oznacza bardzo duży przyrost danych w szybkim tempie. Nie tylko rozmiar danych może stanowić problem. Istotną kwestią jest szybkość przetwarzania (*velocity*), definiowana jako szybkość napływu danych strumieniowych. Tego typu problem wymaga wykorzystania platform o bardzo dużych mocach obliczeniowych. Do rozwiązania tego problemu przydatne są np. chmury obliczeniowe albo wysoko wydajne farmy serwerów. W tym przypadku nie można zastosować relacyjnych baz danych czy hurtowni danych. Różnorodność (*variety*) oznacza natomiast, że mogą być zbierane różne dane, zapisane również w formatach o nieokreślonej strukturze. Jedynie około 10% danych to dane przechowywane w formie uporządkowanych tabel, które można analizować popularnymi narzędziami baz danych. Pozostałe 90% dostępnych danych to dane nieustrukturyzowane, takie jak nagrania z monitoringów, sygnały GPS, zdjęcia satelitarne, ale również dane o temperaturze, przepływach czy częstotliwościach – innymi słowy dane pochodzące z różnego rodzaju urządzeń czy maszyn.

Uzupełnioną definicję prezentuje firma IBM, która określa dodatkowy atrybut :wiarygodność (*veracity*), a model ten można oznaczyć jako 4V. Wiarygodność danych jest istotna z punktu widzenia analizy danych. Jest to również ważna w przypadku analizy danych mających wpływ np. na proces projektowania.

Z kolei pojęcie *Biga Data* według Instytutu SAS jest definiowane jako duża ilość danych, strukturalnych i niestrukturalnych, które ciągle napływają [3]. Pojęcie jest opisywane według atrybutów takich, jak objętość (*volume*), szybkość przetwarzania (*velocity*), różnorodność (*variety*). Dodatkowo Instytut SAS zwraca uwagę na zmienność (*variability*) i złożoność (*complexity*).

Warto podkreślić, że już teraz analiza dużych zbiorów danych może stanowić podstawę konkurencji i przyczyniać się do zwiększenia innowacyjności wytwarzanych produktów. Należy pamiętać o tym, że wykorzystanie technologii *Big Data* umożliwia wyciąganie głębszych i istotniejszych wniosków niż powszechnie stosowana do tej pory tradycyjna analiza danych [4, 5].

### Wybrana metoda przetwarzania danych w technologii *Big Data*

Rozwój technologii IT, a w szczególności dynamiczne zwiększanie ilości i szybkości napływu danych, wymaga zmiany sposobu przetwarzania informacji. Wiąże się to z inwestycjami w infrastrukturę informatyczną (zakup serwerów, pamięci, macierzy danych). Dodatkowo należy zastosować odpowiednie, dedykowane do przetwarzania dużych ilości danych, oprogramowanie. Często wykorzystywanymi rozwiązaniami, dedykowanymi do zastosowań *Big Data* są np. model *Cloud Computing* [6], projekt *Stratosphere* [7] oraz *Apache Hadoop* [8]. Zdecydowanie najpopularniejsze wśród wymienionych to *Apache Hadoop*. Jest on zestawem oprogramowania (zbiorem bibliotek), umożliwiającym pracę w technologii *Big Data*. Framework ten początkowo był rozwijany przez *Apache Foundation*. Jest to rozwiązanie typu *open-source*. Obecnie *Apache Hadoop* jest pewnego rodzaju standardem przechowywania i przetwarzania wielu terabajtów, a nawet petabajtów danych. Architektura *Apache Hadoop* składa się z kilku modułów:

- *Hadoop Common* – zawiera biblioteki i narzędzia udostępniane innym modułom,

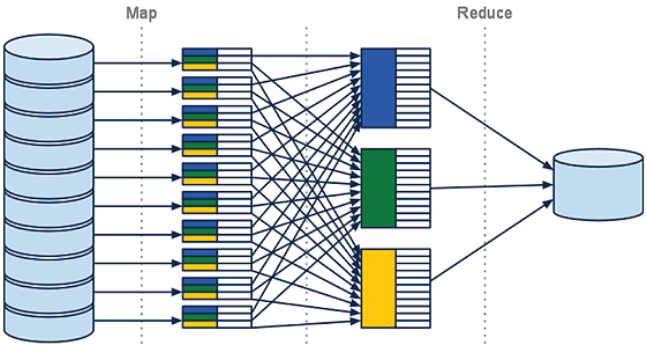
* Dr inż. Rafał Raczko (raczko@sgh.waw.pl) – Szkoła Główna Handlowa w Warszawie

- *Hadoop Distributed File System* – rozproszony system plików, w którym przechowywane są dane, zapewniając wysoką przepustowość,
- *Hadoop YARN* – system służący do zarządzania zasobami danych w klastrach oraz harmonogramowania zadań użytkowników,
- *Hadoop MapReduce* – model programistyczny zaprojektowany do przetwarzania dużej ilości danych.

*Apache Hadoop* jest wykorzystywany do przetwarzania bardzo dużych ilości danych, przesyłanych w sposób ciągły i stały, oraz gdy jest potrzeba ich dystrybucji do innych węzłów. W klasycznym podejściu nie ma możliwości bieżącego obsługiwania terabajtów danych, które napływają i powiększają się w bardzo szybkim tempie. Dane przetwarzane są na jednym superkomputerze (rys. 2). Rozwiązanie takie jest nieefektywne, a bezpieczeństwo danych wątpliwe.

*Apache Hadoop* umożliwia natomiast podzielenie danych na mniejsze części i ich równoległe przetwarzanie z wykorzystaniem wielu różnych węzłów. Dodatkowo należy podkreślić, że zapewniona jest tu możliwość obsługiwania danych niestrukturalnych. Każda z części danych jest przetwarzana przez niezależne komputery. Wzrost ilości danych wymusza jedynie dodanie kolejnego węzła *Apache Hadoop*. Rozproszony system danych daje jeszcze jedną korzyść. Poszczególne serwery w klastrze zdecydowanie nie muszą już być wysoko wydajnymi superkomputerami. Inną zaletą jest większe bezpieczeństwo danych z uwagi na fakt, że każdy przechowywany plik jest kilkakrotnie replikowany. Awaria jednego węzła nie powoduje utraty danych.

Po przetworzeniu danych należy je wszystkie zintegrować, ponieważ każdy z węzłów wykonuje jedynie część algorytmu. Ostatecznym wynikiem jest więc połączenie wszystkich danych w całość z wykorzystaniem modelu *MapReduce*, który składa się z dwóch głównych funkcji *Map* oraz *Reduce*. Schemat działania algorytmu *MapReduce* przedstawiono na rys. 2.




Rys. 2. *Apache Hadoop MapReduce* – schemat działania [9]

*Map* jest to funkcja mapująca. Przyjmuje jako parametr listę parametrów wejściowych, a następnie generuje kolejną listę zawierającą zmapowane wartości. W przypadku oprogramowania *Apache Hadoop*, ma to na celu podział danych na mniejsze części, które następnie z łatwością mogą być przetwarzane całkowicie niezależnie.

*Reduce* jest funkcją redukującą. Wszystkie dane, które posiadają ten sam klucz wygenerowany wcześniej przez funkcję mapującą, trafiają razem do tego samego węzła redukującego. Zadaniem funkcji redukującej jest integracja danych. Wyjście ze wszystkich węzłów redukujących to wynik algorytmu.

## Wykorzystanie *Big Data*

Rozważając obszary wykorzystujące *Big Data* warto zauważyć, że zakres zastosowań jest niezwykle szeroki. Rozwiązania *Big Data* mogą być z powodzeniem stosowane w inżynierii w zakresie projektowania, wytwarzania czy rozwoju produktów, gdzie zakres przetwarzanych i analizowanych danych często jest bardzo duży.

W procesie projektowania innowacyjnego produktu inżynier musi przetworzyć ogromne ilości danych pochodzące z wielu różnych źródeł. Często poddawane są analizie dane historyczne, z zakresu życia danego produktu, jak również informacje od klienta.

Również w procesach produkcyjnych technologia *Big Data* może być z powodzeniem wykorzystywana. Chcąc przewidzieć pojawiające się awarie lub zoptymalizować proces produkcyjny, stosuje się różnego rodzaju czujniki i detektory, które generują duże ilości danych. Przetworzenie i analiza tych danych mogą wspierać proces produkcyjny.

Dzięki możliwości przetwarzania dużych ilości danych o luźnej strukturze jest także możliwość modelowania zjawisk pogodowych w celu optymalizacji rozmieszczenia turbin wiatrowych, jak również maksymalizacji generowanej mocy i czasu życia turbiny.

Technologia *Big Data* jest coraz częściej stosowana np. w branży motoryzacyjnej. W przeciągu kilku lat możemy doświadczyć prawdziwego przełomu w tej dziedzinie, zarówno w bezpieczeństwie, jak i funkcjonalności pojazdów. Jest to możliwe również dzięki analizie danych. Dane te mogą być pomocne w procesie tworzenia pojazdów, które będą w stanie jeździć automatycznie – bez potrzeby sterowania przez kierowcę. Automatyczne pojazdy miałyby nie tylko poprawić bezpieczeństwo na drogach, ale także zoptymalizować odległości między jadącymi pojazdami. Dzięki łączeniu danych z map z danymi wytwarzanymi i przetwarzanymi w czasie rzeczywistym pochodzącymi z czujników samochodów, auta mogłyby podejmować decyzje o zachowaniu na drogach.

Przykładem wykorzystania *Big Data* w branży motoryzacyjnej jest nowoczesny samochód hybrydowy marki Ford (Ford Fusion). Jego system komputerowy gromadzi i przetwarza więcej niż 25 gigabajtów danych na godzinę. Wśród gromadzonych informacji znajdują się dane dotyczące bezpośrednio samochodu (takie jak jego przyspieszenie, odchylenie od kursu, szybkość), dane związane z zachowaniem kierowcy oraz dane biometryczne kierowcy (m.in. częstotliwość oddechów, tętno). Całościowy obraz sytuacji uzyskany na podstawie analizy tych danych umożliwia lepsze zrozumienie zachowania kierowców na drogach, a także przekłada się na zmniejszenie liczby wypadków.

## Podsumowanie

Wykorzystanie rozwiązań *Big Data* stwarza bardzo wiele nowych możliwości w inżynierii, których wcześniej nikt nie dostrzegał. Jest to osiągalne dzięki specyfice *Big Data*, a więc możliwości przetwarzania i analizy dużych ilości szybko napływających i złożonych danych. W Polsce technologia *Big Data* stopniowo zyskuje na popularności, a przedsiębiorstwa, które szybko zauważą jej szerokie możliwości, mogą osiągnąć dużą przewagę konkurencyjną.

LITERATURA

1. Witten H., Frank E., Hall M. „*Data Mining: Practical Machine Learning Tools and Techniques*". Morgan Kaufmann, 2011.
2. Chardonnens T. „Big Data analytics on high velocity streams". *Software Engineering Group Department of Informatics University of Fribourg*, s. 7.
3. Instytut SAS, http://www.sas.com.
4. Manyika J., Chui M., Brown B., Bughin J., Dobbs R., Roxburgh C., Hung Byers A. „Big Data: The next frontier for innovation, competition, and productivity". *McKinsey Global Institute*, 2011.
5. Guterman J. „*Release 2.0: Issue 11*". O'Reilly Media, 2009.
6. Jadeja Y., Modi K., „Cloud computing – concepts, architecture and challenges", *International Conference on Computing. Electronics and Electrical Technologies (ICCEET)*, IEEE. India, 2012.
7. Projekt Stratosphere, http://stratosphere.eu.
8. Apache Hadoop, http://hadoop.apache.org.
9. Artificial Intelligence in Motion, http://aimotion.blogspot.com/2012/08/introduction-to-recommendations-with.html.