Duże zbiory danych
big data

dr hab. inż. Przemysław Korytkowski

Wykład 1

1

Prowadzący

- dr hab. inż. Przemysław Korytkowski, pkorytkowski@zut.edu.pl
- dr inż. Bartłomiej Małachowski, bmalachowski@zut.edu.pl
- dr hab. inż. Jarosław Jankowski, prof. ZUT, jjankowski@zut.edu.pl

2

Literatura

- Google
- apache.org
- White, Tom (2016) Hadoop. Kompletny przewodnik, Helion, Gliwice.

3

Treści programowe

- Czym jest Big data?
- Hadoop: HDFS i YARN
- Hive
- Spark
- Sieci złożone

4

Zasady zaliczenia

- Test na koniec semestru z wykładów – 50%.
- Ocena z laboratoriów – 50%

5

The World's Technological Capacity to Store, Communicate, and Compute Information

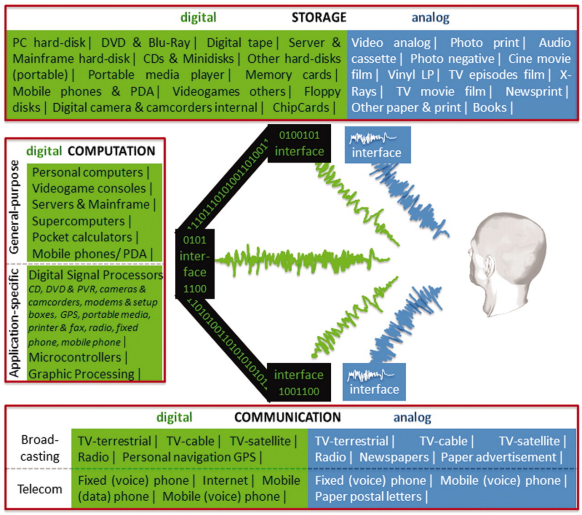
by Martin Hilbert, and Priscila López

Science
Volume 332(6025):60-65
April 1, 2011

Science
AAAS

6

Fig. 1 The three basic information operations and their most prominent technologies.

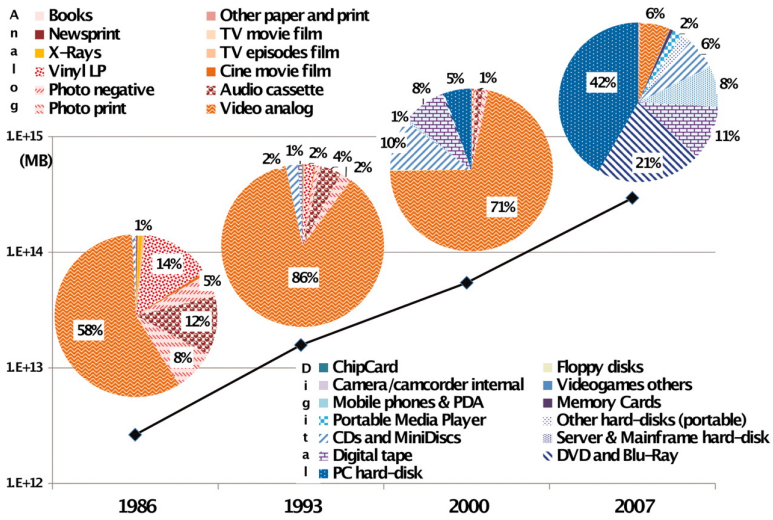


Martin Hilbert, and Priscila López Science 2011;332:60-65

Science AAAS

7

Fig. 2 World's technological installed capacity to store information (table SA1) (16).

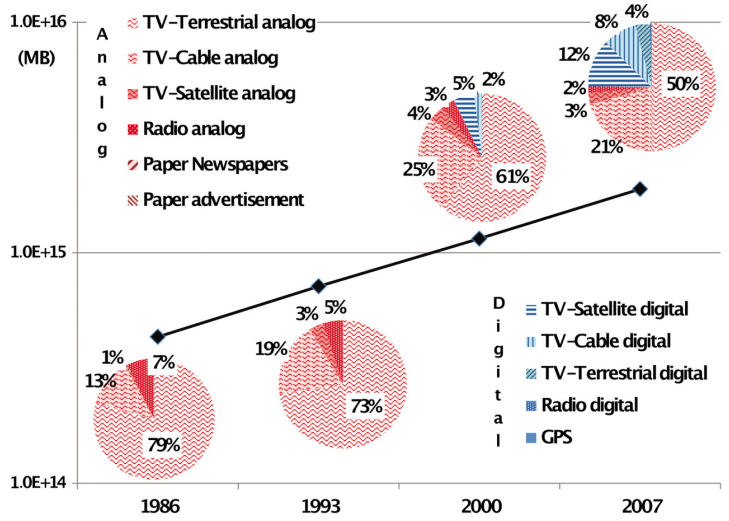


Martin Hilbert, and Priscila López Science 2011;332:60-65

Science AAAS

8

Fig. 3 World's technological effective capacity to broadcast information in optimally compressed megabytes MB per year, for 1986, 1993, 2000, and 2007; semi-logarithmic plot (table SA2) (16).

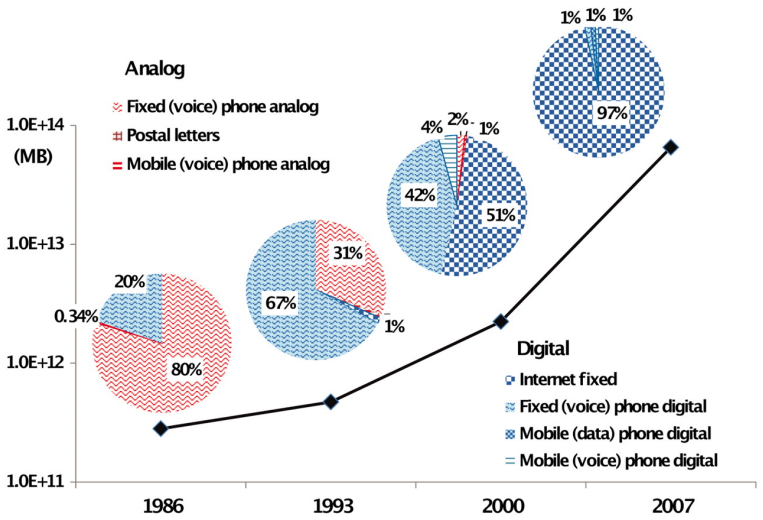


Martin Hilbert, and Priscila López Science 2011;332:60-65

Science AAAS

9

Fig. 4 World's technological effective capacity to telecommunicate information (table SA2) (16).

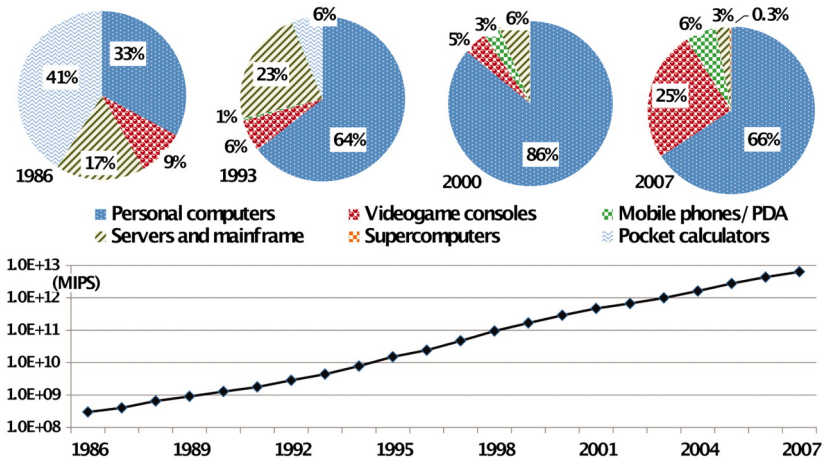


Martin Hilbert, and Priscila López Science 2011;332:60-65

Science AAAS

10

Fig. 5 World's technological installed capacity to compute information on general-purpose computers, in MIPS (table SA3) (16).

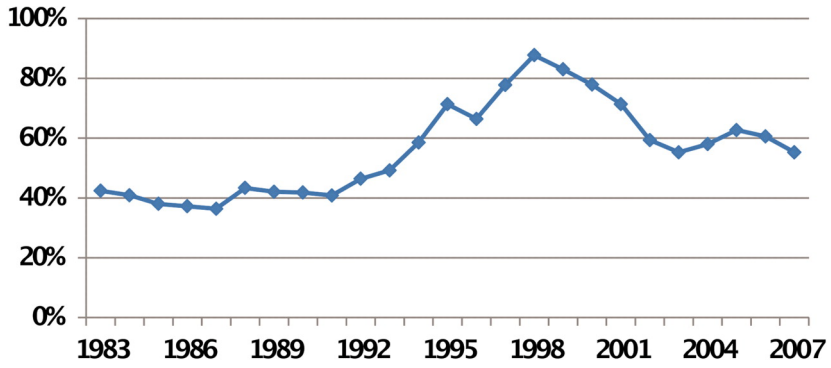


Martin Hilbert, and Priscila López Science 2011;332:60-65

Science AAAS

11

Fig. 6 Annual growth of installed general-purpose computational capacity as percentage of all previous computations since 1977 (year t / Σ[1977, year t – 1]) (table SA2) (16).

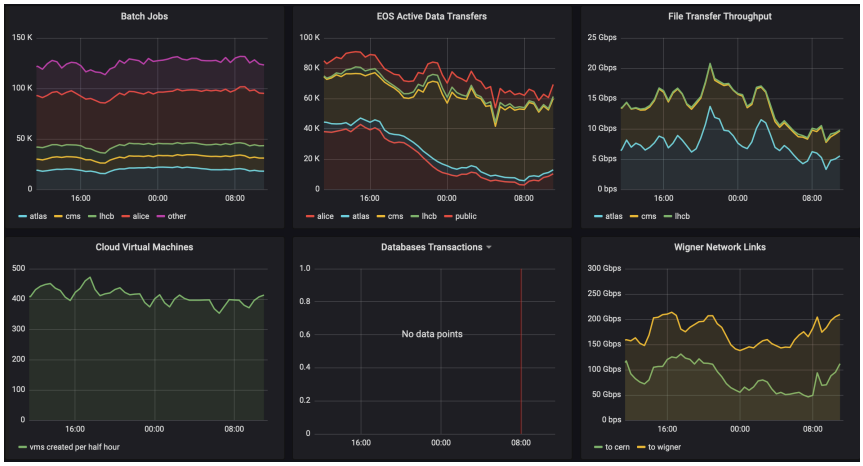


Martin Hilbert, and Priscila López Science 2011;332:60-65

Science AAAS

12

CERN data center

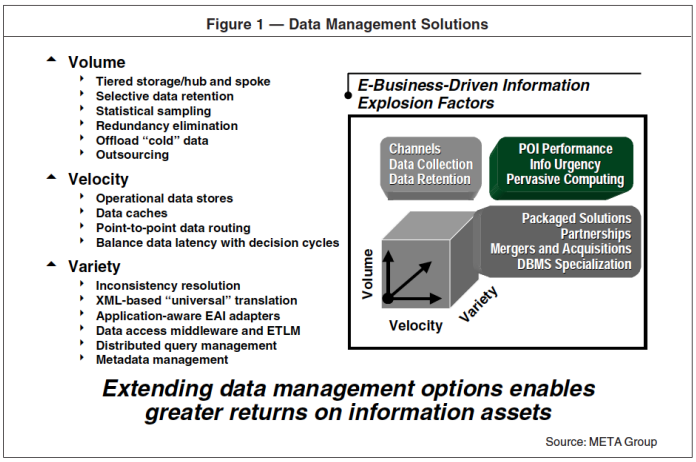


https://home.cern/science/computing/data-centre

13

Big data – początki

- 6 luty 2001
- **Doug Laney**
- META Group
- 3D Data Management



14

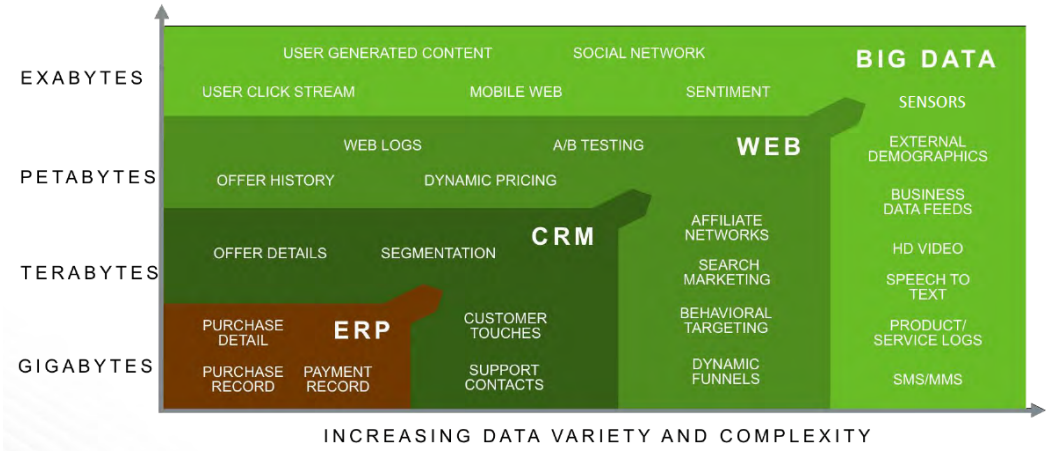
Big data

- **Ilość (volume).** Organizacje zbierają dane z różnorodnych źródeł: transakcje biznesowe, media społecznościowe, dane z sensorów, dane wymieniane między urządzeniami. W przeszłości przechowywanie tych danych stanowiło problem, ale obecnie nowe technologie (takie jak np Hadoop) znacznie to ułatwiły.
- **Szybkość (velocity).** Dane powstają i są dostarczane niezwykle szybko i muszą być obsługiwane z odpowiednim reżimem czasowym. Znaczniki RFID, czujniki i inteligentne mierniki wymagają obsługi olbrzymich ilości danych w czasie zbliżonym do rzeczywistego.
- **Różnorodność (veriety).** Dane przychodzą w różnych formatach – od ustrukturyzowanych, numerycznych danych w tradycyjnych bazach danych do niestrukturalnych dokumentów tekstowych, email, video, audio, danych znaczników magazynowych lub transakcji finansowych.
- **Zmienność (variability).** Poza coraz większym tempem napływu danych i ich rosnącą różnorodnością przepływy danych mogą podlegać dużym wahaniom okresowym. Czy w mediach społecznościowych jakiś temat jest szczególnie popularny? Czasami trudno jest zarządzać danymi napływającymi w trakcie szczytu dziennego, sezonowego czy wywołanego konkretnym wydarzeniem. Jest to jeszcze trudniejsze w przypadku danych niestrukturalnych.
- **Złożoność (complexity).** Dane napływają do nas z wielu różnych źródeł. Łączenie, dopasowywanie, oczyszczanie i przekształcanie danych w różnych systemach to zadania wymagające dużego wysiłku. Niemniej jednak łączenie i zestawianie relacji, hierarchii i różnorodnych powiązań między danymi jest niezbędne. W przeciwnym razie potoki danych mogą łatwo wymknąć się spod kontroli.

Źródło: www.sas.com

15

3V = Volume + Velocity + Variety



Źródło: www.cloudera.com

16

Dlaczego big data są ważne?

Wartość big data nie zależy od tego ile mamy danych, ale od tego w jaki sposób je wykorzystamy. Przykładowo, możemy wykorzystać dane z wielu źródeł aby:

1. zmniejszyć koszty,
2. zredukować czas,
3. wytworzyć nową ofertę produktową,
4. podjąć lepsze decyzje.

Jeśli połączymy dane masowe z zaawansowaną analityką, możemy wspomóc operacje biznesowe, takie jak:

- Określenie przyczyn awarii, nieprawidłowej pracy, defektów w czasie zbliżonym do rzeczywistego.
- Generowanie kuponów w miejscu zakupów, bazując na zwyczajach zakupowych klienta.
- Przeliczanie ryzyka całego portfela w ciągu minut.
- Wykrywanie zachowania wskazującego na nadużycie, zanim znacząco wpłynie ono na organizację.

Źródło: www.sas.com

17

Kto używa big data?

Banki

W bankowości strumienie danych napływających z wielu źródeł dają możliwość odkrywania nowej wiedzy i innowacyjnych sposobów zarządzania danymi masowymi. Z jednej strony kluczowe jest zrozumienie klienta i podniesienie jego satysfakcji z oferowanych mu usług, ale z drugiej strony konieczne jest minimalizowanie ryzyka i redukcja potencjalnych nadużyć oraz zapewnienie zgodności z regulacjami instytucji nadzorczych. Big data pozwala na uzyskanie kompleksowej wiedzy, ale tylko wtedy, gdy instytucje finansowe przejdą na wyższy poziom wykorzystania zaawansowanej analityki.

Źródło: www.sas.com

18

Kto używa big data?

Produkcja

Przedsiębiorstwa wyposażenie w informacje pochodzące z analizy dużych zbiorów danych mogą podnieść jakość produktów, zwiększyć wydajność produkcji oraz ograniczyć straty – co ma kluczowe znaczenie do osiągnięcia sukcesu na obecnym bardzo konkurencyjnym rynku. Coraz więcej producentów pracuje w trybie kultury analitycznej przez co mogą szybciej rozwiązywać problemy i podejmować trafne decyzje biznesowe.

Źródło: www.sas.com

19

Kto używa big data?

Handel detaliczny

Budowanie trwałych relacji z klientami ma ogromne znaczenie dla rozwoju w branży handlu detalicznego, a jednym ze sposobów osiągnięcia tego celu jest odpowiednie zarządzanie big data. Handlowcy potrzebują optymalnych sposobów dotarcia do klientów, najbardziej efektywnych sposobów na zarządzanie transakcjami oraz, co jest szczególnie istotne z punktu widzenia strategii, sposobów na odzyskanie utraconych szans sprzedażowych. Big data pozostaje w centrum wszystkich tych działań.

Źródło: www.sas.com

20

Kto używa big data?

Ochrona zdrowia

W przypadku ochrony zdrowia wszystko musi być robione szybko i dokładnie, a także bardzo często z zachowaniem reguł transparentności i bezpieczeństwa wymaganego przez szczegółowe regulacje. Efektywne zarządzanie big data pozwala służbie zdrowia odkryć nieznane zależności oraz polepszyć obsługę pacjenta.

Źródło: www.sas.com

21

Kto używa big data?

Sektor publiczny

Instytucje publiczne mogą wykorzystać analitykę bazującą na gromadzonych danych masowych, aby usprawnić zarzadzanie, optymalizować koszty, zwiększać jakości obsługi obywateli oraz przeciwdziałać przestępczości. Oczywiście istotne znaczenie mają tu wymogi związane z transparentnością działań oraz ochroną prywatności obywateli.

Źródło: www.sas.com

22

Kto używa big data?

Edukacja

Osoby odpowiedzialne za kształcenie, dzięki wiedzy pochodzącej z analizy wielkich zbiorów danych mogą wnieść istotny wkład w rozwój systemu oświaty i programów nauczania. Dzięki analizie big data można identyfikować zagrożenia dla uczniów, pomagać studentom w wyborze właściwej ścieżki edukacji oraz usprawnić system oceny i wsparcia nauczycieli.

Źródło: www.sas.com

23

Wartość danych

Należy pamiętać, że główna wartość big data nie pochodzi z danych w ich surowej postaci, ale z wyników ich przetworzenia i analizowania prowadzących do wniosków, produktów i usług, które są pochodną analiz.

Rewolucyjne zmiany w technologiach big data i podejściach do zarządzania wymagają równie rewolucyjnych zmian w zakresie wykorzystania danych w organizacji do wsparcia podejmowania decyzji oraz rozwoju innowacyjnych produktów i usług.

Źródło: www.sas.com

24

Źródła danych

- **Dane strumieniowe**

  Kategoria ta zawiera dane napływające do systemów informatycznych z Internetu lub podłączonych urządzeń. Możliwe jest analizowanie takich danych w ruchu – w momencie, gdy napływają oraz podejmowanie decyzji, które dane należy przechowywać, które nie są istotne oraz które wymagają dalszych analiz.
- **Dane z mediów społecznościowych**

  Dane takie są coraz bardziej atrakcyjnym zbiorem informacji, szczególnie dla zastosowań w marketingu, sprzedaży i obsłudze klienta. Często są one nieustrukturyzowane lub tylko częściowo ustrukturyzowane, przez co wyzwaniem staje się wykorzystanie ich w zastosowaniach analitycznych.
- **Dane dostępne publicznie**

  Wielkie zbiory danych dostępne są również z publicznych źródeł takich jak np. organizacje rządowe lub agendy Unii Europejskiej.
- **Dane z wewnętrznych baz danych**

Źródło: www.sas.com

25

Wykorzystanie danych

W jaki sposób składować i zarządzać danymi?

Przechowywanie danych mogło być problemem kilka lat temu, ale obecnie dostępne są relatywnie tanie opcje rozwiązań, które mogą być najlepszą strategią w tym zakresie dla każdej organizacji.

Jak wiele danych analizować?

Wiele organizacji nie wyklucza żadnych danych ze swoich analiz, co jest obecnie możliwe dzięki wykorzystaniu technologii high-performance computing takich jak np. przetwarzanie grid lub analityka in-memory. Innym podejściem jest określenie z góry, jeszcze przed wykonaniem analiz, jaki podzbiór danych ma znaczenie.

Jak wykorzystać wyniki analiz?

Im szerszą posiadamy wiedzę, z tym większym zaufaniem możemy podejmować decyzje biznesowe. Rozsądnym podejściem jest zbudowanie strategii bazującej na posiadanych informacjach.

Źródło: www.sas.com

26

Wybór technologii

- Tanie i pojemne przechowywanie (storage)
- Szybkie procesory
- Dostępne platformy open source, takie jak Hadoop
- Przetwarzanie równoległe, klastry obliczeniowe, MPP, wirtualizację, duże środowiska grid, szybkość połączenia i transferu
- Przetwarzanie w chmurze i inne opcje architektoniczne

Źródło: www.sas.com

27

Dziękuję za uwagę!

28