Wrocław
University
of Science
and Technology

Przetwarzenie danych masowych

Wykład 1 – Wprowadzenie do problemu przetwarzania danych masowych. Zasady zaliczenia.

dr hab. inż. Tomasz Kajdanowicz, Piotr Bielak, Roman Bartusiak

October 12, 2021

HR EXCELLENCE IN RESEARCH

Overview

- Organizacja przedmiotu
  - Prowadzący
  - Zajęcia
  - Syllabus wykładu
  - Syllabus wykładu
  - Ocenianie
  - Materiały
- Wprowadzenie do Przetwarzania Danych Masowych

Overview

- Organizacja przedmiotu
  - Prowadzący
  - Zajęcia
  - Syllabus wykładu
  - Syllabus wykładu
  - Ocenianie
  - Materiały

Wprowadzenie do Przetwarzania Danych Masowych

Prowadzący

Organizacja przedmiotu

- dr hab inż. Tomasz Kajdanowicz
  `tomasz.kajdanowicz@pwr.edu.pl`
- Roman Bartusiak
  `roman.bartusiak@pwr.edu.pl`
- Krzysztof Rajda
  `krzysztof.rajda@pwr.edu.pl`

Godziny konsultacji zostaną przekazane w najbliższym czasie, ale już teraz zapraszamy do 441 A1. Proszę o przesłanie maila i umówienie się na spotkanie.

Zajęcia

Organizacja przedmiotu

Wykład

- Wprowadzenie teoretyczne
- Przegląd podstawowych zagadnień
- Wykłady i laboratoria wzajemnie się uzupełniają

Laboratoria

- Projekt wiodący przez laboratoria
- możliwość wykorzystania AWS
- Terminowość
- Kodowanie

Syllabus wykładu

Organizacja przedmiotu

1. *Wprowadzenie do problemu przetwarzania danych masowych. Zasady zaliczenia.*
2. *Taksonomia metod przetwarzania danych masowych*
3. *Podstawowe metody zrównoleglania algorytmów uczenia maszynowego. Przetwarzanie synchroniczne i asynchroniczne*
4. *Spark - przetwarzanie danych z wykorzystaniem paradygmatu Map-reduce - przetwarzanie wsadowe*
5. *Spark - przetwarzanie danych z wykorzystaniem paradygmatu Map-reduce - przetwarzanie strumieniowe*
6. *Flink - przetwarzanie danych w sposób strumieniowy*
7. *Flink - przetwarzanie danych w sposób wsadowy*

Syllabus wykładu

Organizacja przedmiotu

8. *Produkcyjne aspekty utrzymywania i wdrażania aplikacjami*
9. *Platformy zarządzania zasobami obliczeniowymi - wprowadzenie, OpenStack*
10. *Platformy zarządzania zasobami obliczeniowymi - Kubernetes*
11. *Metody automatyzacji zarządzania produkcyjnymi aplikacjami*
12. *Języki do przetwarzania danych masowych*
13. *Przykładowe metody z rodziny Gradient Boosting Machine*
14. *Zaawansowane metody zrównoleglania algorytmów uczenia maszynowego*
15. *Recap - podsumowanie wykładu*

Ocenianie

Organizacja przedmiotu

Wykład

- Egzamin 7 i 14 lutego 2022
- Ocena z laboratoriów nie ma związku

Laboratoria

- Projekt wiodący przez całość zajęć realizowany w częściach
- Każda część oceniana osobna
- Każda część musi być zaliczona
- Wszystkie części wpływają na ocenę końcową
- Ocena 5.5 do otrzymania za dodatkową pracę

Materiały

Organizacja przedmiotu

- https://lsdp.ml
- AWS Educate – https://awseducate.com
- Github Classroom – https://classroom.github.com
- http://web.stanford.edu/class/cs246/

Overview

- Organizacja przedmiotu
  - Prowadzący
  - Zajęcia
  - Syllabus wykładu
  - Syllabus wykładu
  - Ocenianie
  - Materiały

Wprowadzenie do Przetwarzania Danych Masowych

Big data

Główni gracze:

- Google
- Facebook
- Youtube
- Instagram
- Wikipedia
- Alibaba

Firmy gromadzą petabajty danych w każdej minucie.
W jakim celu?

Powód #1

- Gdyż mogą to zrobić!
- Pamięć dyskowa tanieje zgodnie z prawem Moor'a

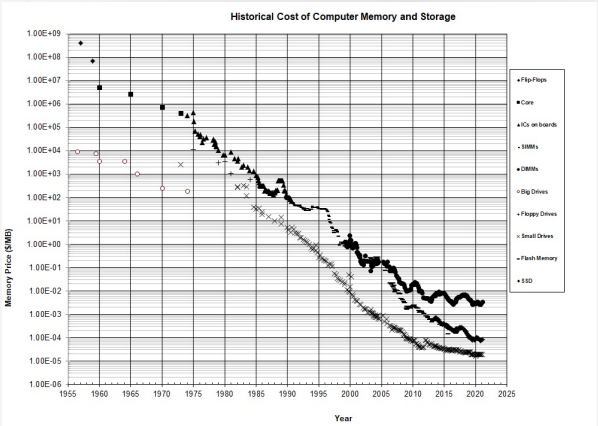


Figure: Źródło: https://jcmit.net/diskprice.htm

Powód #2

- Dane można w prosty sposób monetyzować!
- trend superpersonalizacji (rekomendacje, oferty, promocje, newsfeedy)
- zmiana paradygmatu rynku - marketplace

Ilość danych

- dane to surowiec więcej warty niż ropa
- ilość danych rośnie szybciej niż technologia ich przetwarzania
- pracujący z danymi, zwykle nie są zaangażowani w opracowywanie nowych modeli biznesowych
- rozwijający biznes zwykle nie mają kompleksowej wiedzy na temat całego spektrum dostępnych danych

Figure: Licencja zdjęcia: davidparkins.com, Źródło https://www.economist.com

Ile danych generujemy

Figure: Źródło: https://ec.europa.eu/newsroom/rtd/items/713444/en

Ile danych generujemy

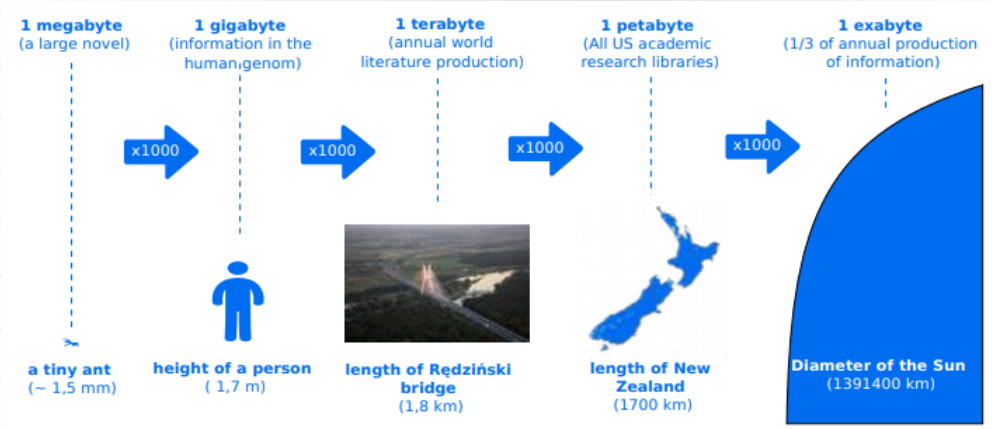


Figure: Źródło: Industry Tap

Przykłady zastosowań big data

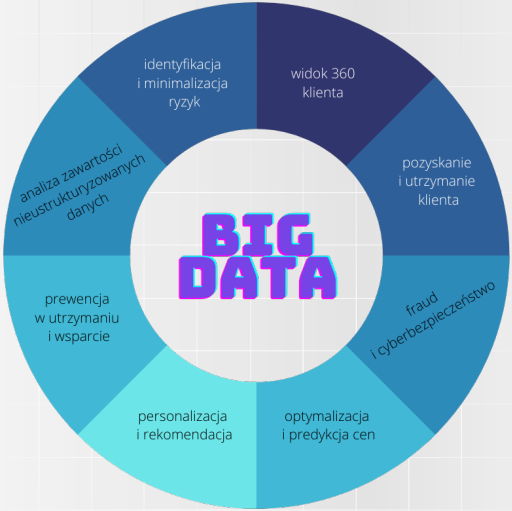
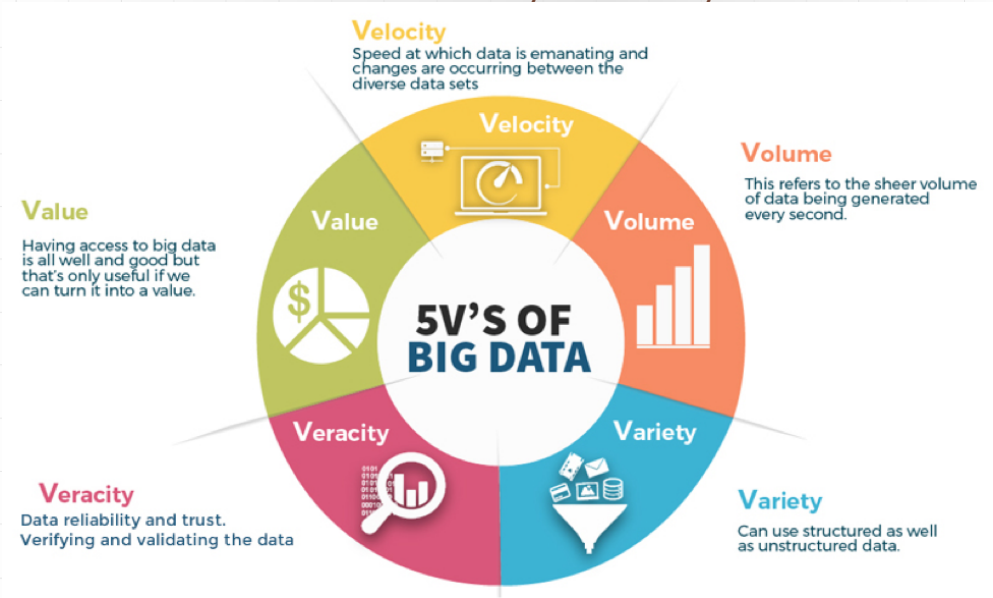


Figure: Źródło: opracowanie własne


Velocity
Speed at which data is emanating and changes are occurring between the diverse data sets
Velocity
Volume
This refers to the sheer volume of data being generated every second.
Volume
Value
Having access to big data is all well and good but that's only useful if we can turn it into a value.
Value
5V'S OF BIG DATA
Veracity
Data reliability and trust. Verifying and validating the data
Veracity
Variety
Can use structured as well as unstructured data.
Variety

Wprowadzenie do Przetwarzania Danych Masowych

- **Volume (wolumen)** – ogromne ilości danych,
- **Velocity (prędkość)** – dane napływają z kilku źródeł z różną i zmienną prękością,
- **Value (wartość)** – dane mogą być trudne w zdobyciu,
- **Veracity (wiarygodność)** – szum, anomalie i przesądy (biases) w danych,
- **Variety (różnorodność)** – wiele źródeł danych, różne typy danych (ustrukturalizowane i nieustrukturalizowane),

Definicja jest czasami rozszerzana do 7V:

- **Validity (poprawność)** – czy dane są poprawne i właściwe dla danego zastosowania,
- **Volatility (ulotność)** – jak długo dane są ważne (valid) oraz jak długo powinny być przechowywane,

Wrocław University of Science and Technology

Przetwarzenie danych masowych

Wykład 1 – Wprowadzenie do problemu przetwarzania danych masowych. Zasady zaliczenia.

dr hab. inż. Tomasz Kajdanowicz, Piotr Bielak, Roman Bartusiak

October 12, 2021