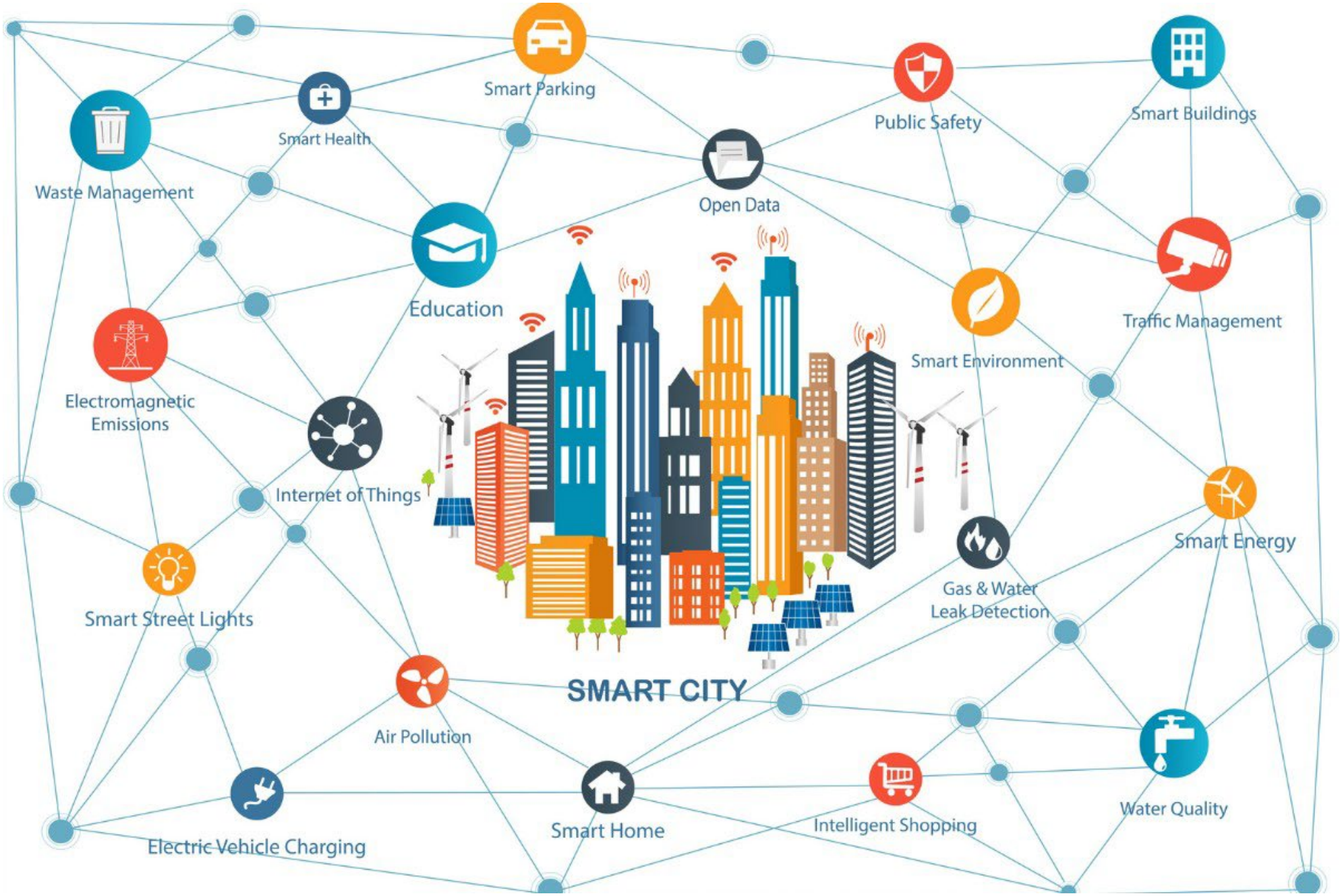
Metody i narzędzia *Big Data*

Krajobraz metod i narzędzi *Big Data*

1956: transport dysku twardego o pojemności 5Mb

1963: gorąca linia nawigacyjna, powstała długo przed usługą Google Maps


Smart Parking
Smart Health
Public Safety
Smart Buildings
Waste Management
Open Data
Education
Traffic Management
Smart Environment
Electromagnetic Emissions
Internet of Things
Smart Energy
Smart Street Lights
Gas & Water Leak Detection
SMART CITY
Air Pollution
Water Quality
Smart Home
Intelligent Shopping
Electric Vehicle Charging

Problem

Dane są zbyt ***duże***, aby dało się je **przechowywać** w klasycznych bazach danych i **przetwarzać** klasycznymi metodami

Archiwa
Skany, oświadczenia, recepty, ...

Dokumenty
PDF, DOC, XLSX, RTF, CSV, HTML, JSON, ...

Aplikacje
E-recepty, Bankowość, gry, ...

Media
Zdjęcia, filmy, audio, zdalne spotkania, ...

Sieci społecznościowe
Facebook, Twitter, LinkedIn, Spotify ...

Internet rzeczy
Klimatyzatory, TV, inteligentny dom, inteligentny magazyn, ...

INTERNET RZECZY

Dane sensoryczne, pomiary
Systemy monitoringu (kamery, czujniki), systemy czasu rzeczywistego, lokalizacja, GPS, Samoloty, Roboty przemysłowe,...

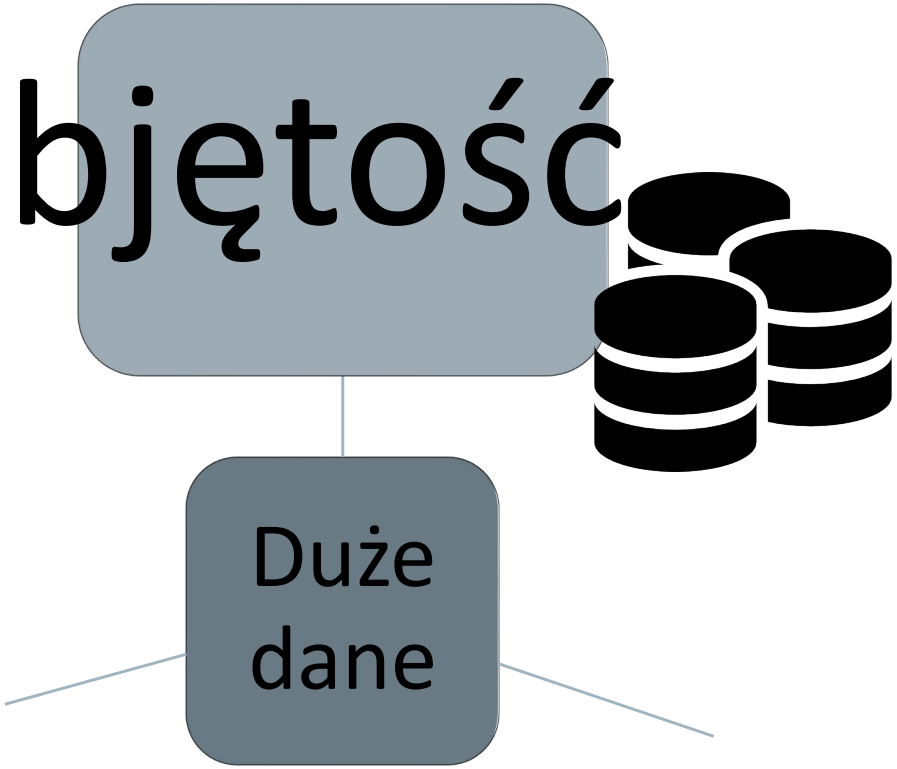
Rejestry aplikacji
Serwery, logi, „kliknięcia” na witrynie,...

Repozytoria
NoSQL, Hadoop, RDBMS ,...

Co to znaczy, że dane są **duże**?

Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." *International journal of information management* 35.2 (2015): 137-144.

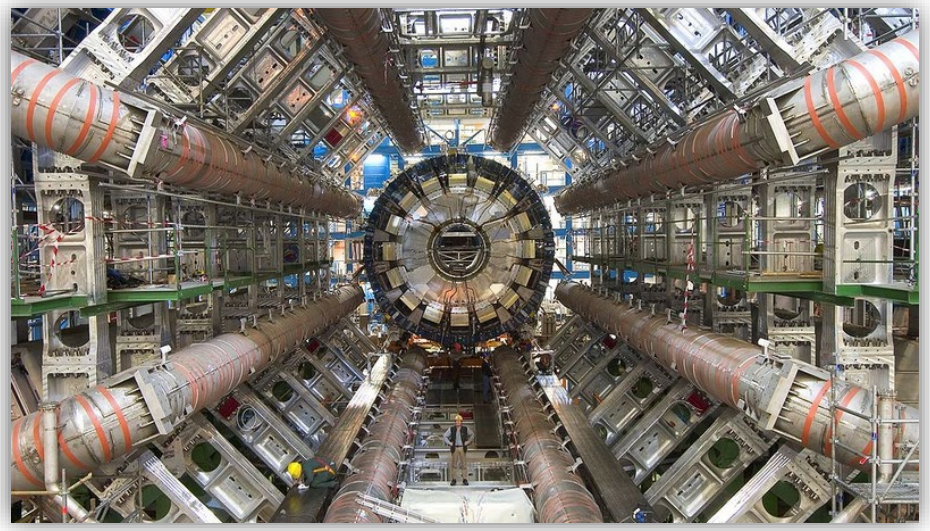
Co to znaczy, że dane są **duże**?

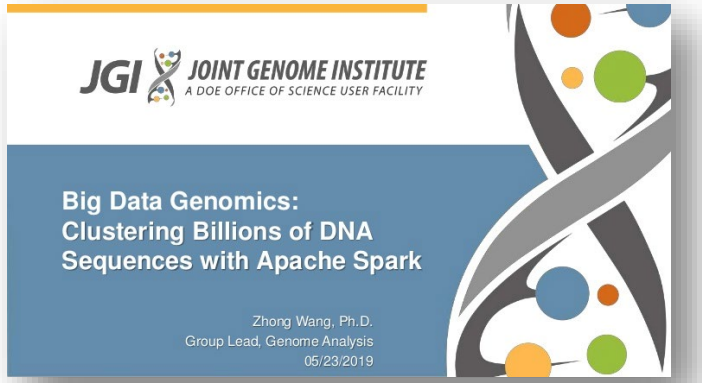


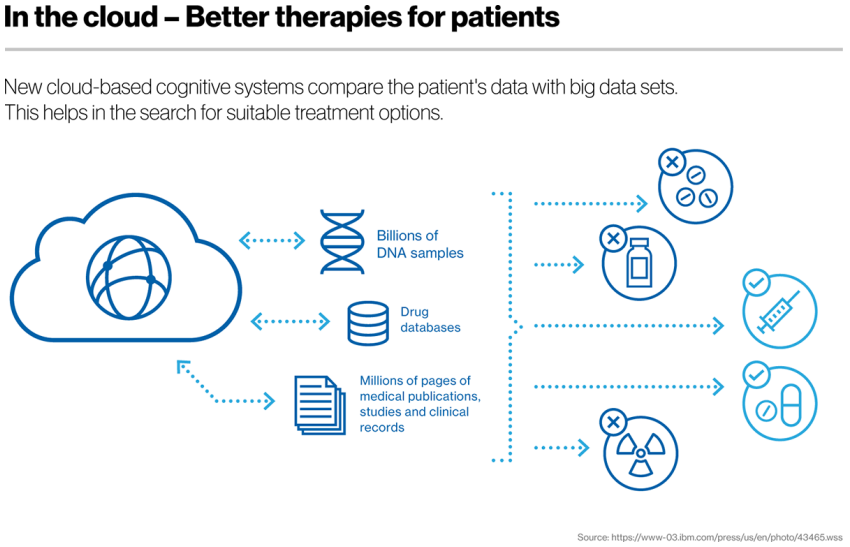
Duża objętość

Przykład

- Objętość liczona w petabajtach
- Wyzwania: przechowywanie i przetwarzanie
  - nie da się ich załadować na pojedynczą maszynę i przetwarzać



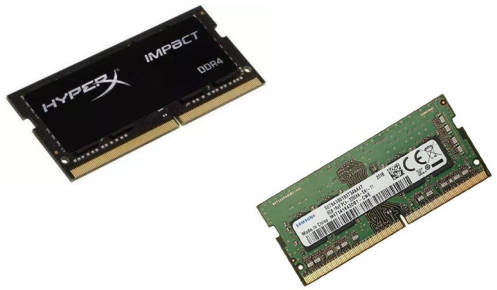




Duża objętość

Przykład

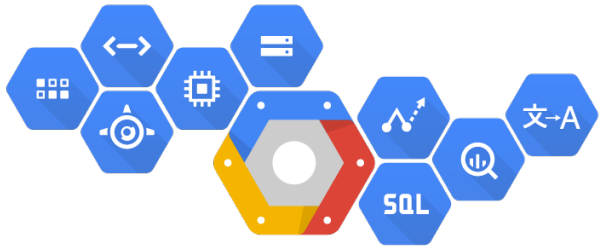
Czy mieszczą się w pamięci?	Czy mieszczą się na dysku?	Jakie to dane?
tak	tak	małe
nie	tak	średnie
nie	**nie**	**duże**



Oussous, Ahmed, et al. "Big Data technologies: A survey." *Journal of King Saud University-Computer and Information Sciences* 30.4 (2018): 431-448.

Duża objętość

Technologie



Google Cloud Platform

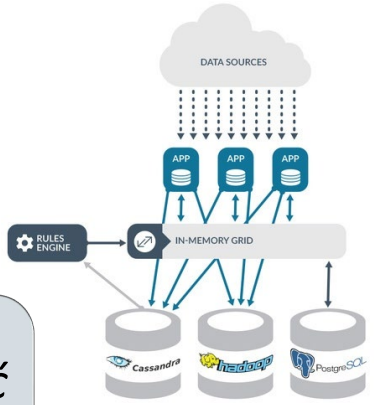
Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." *International journal of information management* 35.2 (2015): 137-144.

Co to znaczy, że dane są **duże**?

Objętość

Duże dane

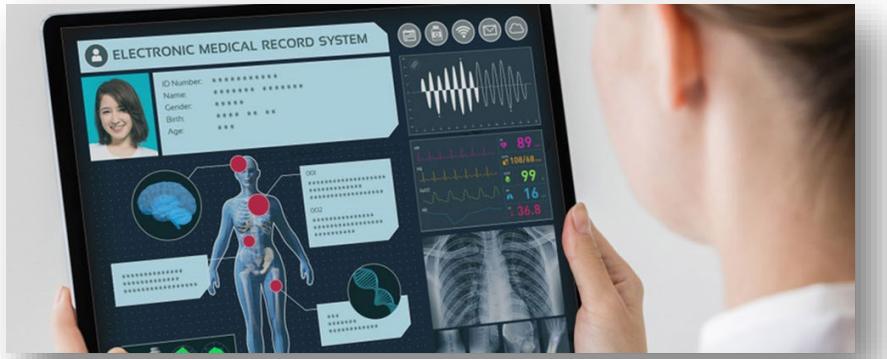
Szybkość i zmienność (strumienie)

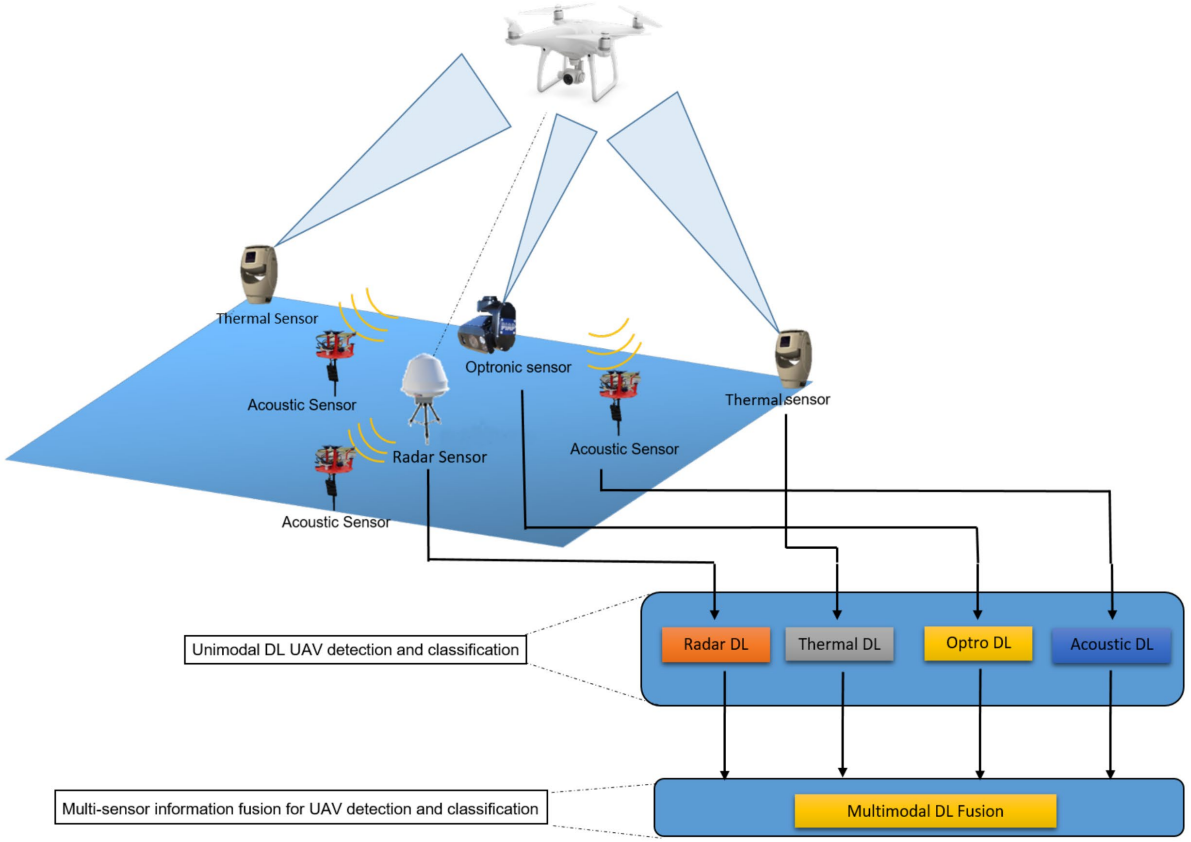


Szybko narastają

Duża szybkość i zmienność (**dane strumieniowe**)
Przykład

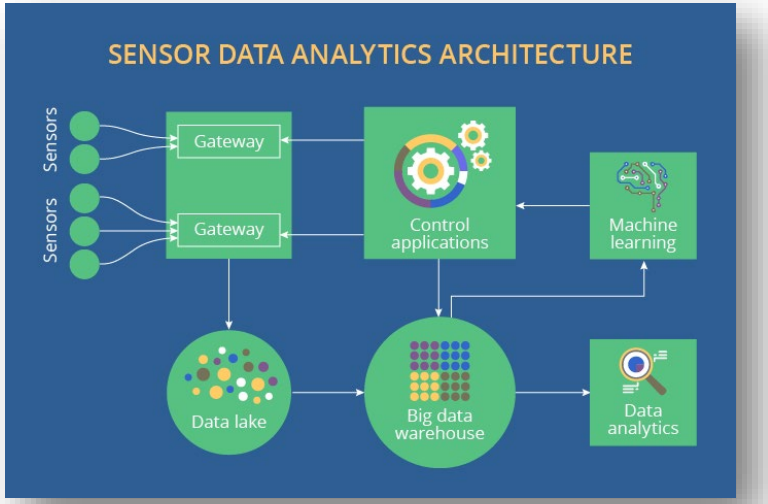
- Wyzwania:
  - analiza na bieżąco: **przetwarzanie w czasie rzeczywistym**, niestacjonarność
  - integracja danych z różnych źródeł
  - integracja z historycznymi danymi
  - niekompletność, zdublowane dane



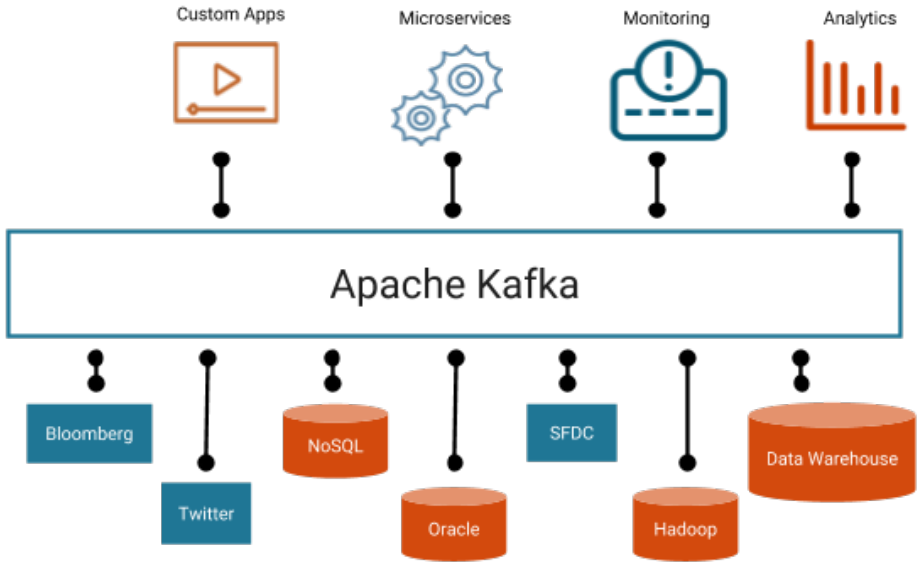


Duża szybkość i zmienność (**dane strumieniowe**)

Technologie



Apache Beam

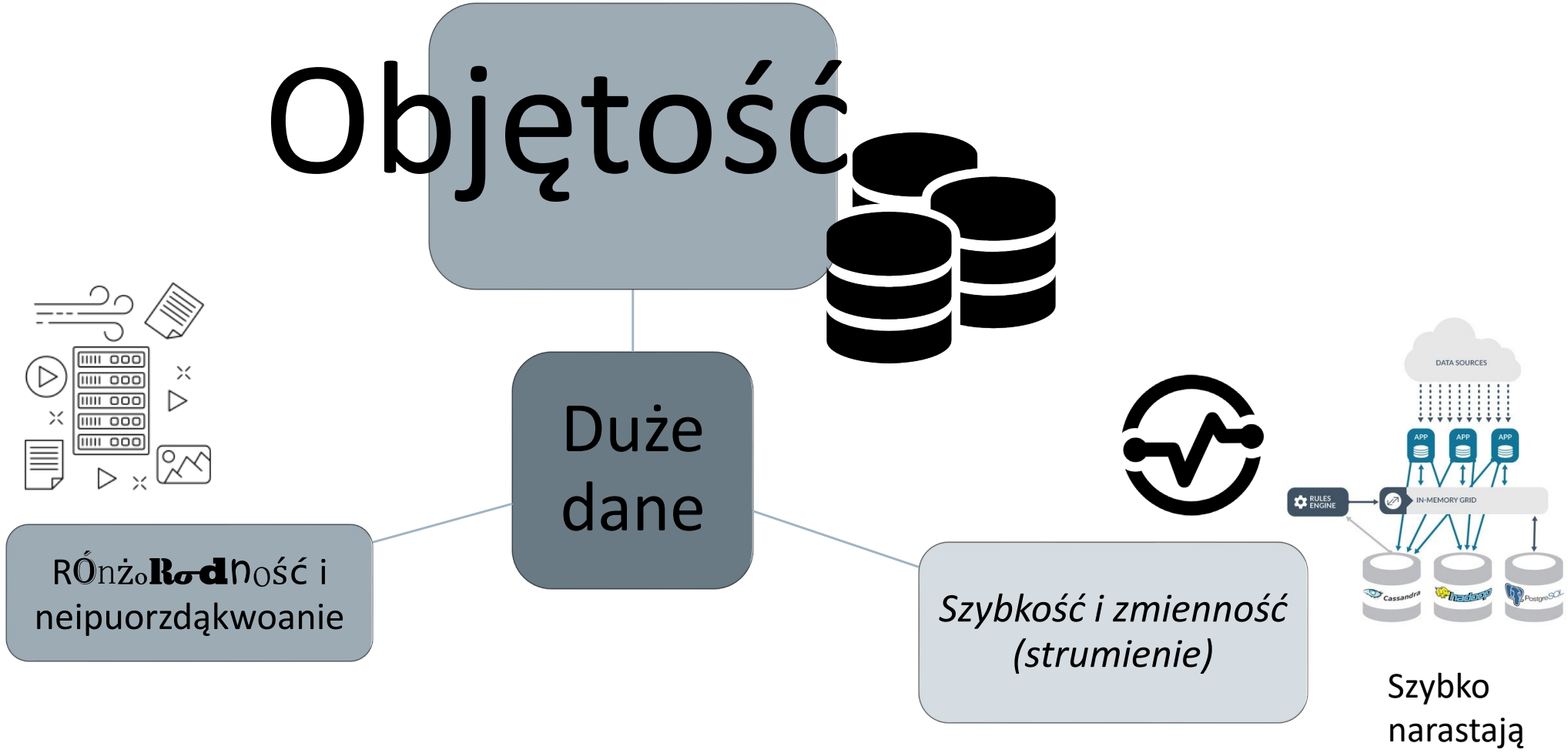


APACHE kafka®

A distributed streaming platform

Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." *International journal of information management* 35.2 (2015): 137-144.

Co to znaczy, że dane są **duże**?



Duży bałagan (ang. *complexity*)

Przykłady

- **Nieustrukturyzowane**, słabo udokumentowane, wielopoziomowe, pochodzące **z różnych źródeł**
- Wyzwania: przekształcanie danych do określonego formatu

Duży bałagan (ang. *complexity*)

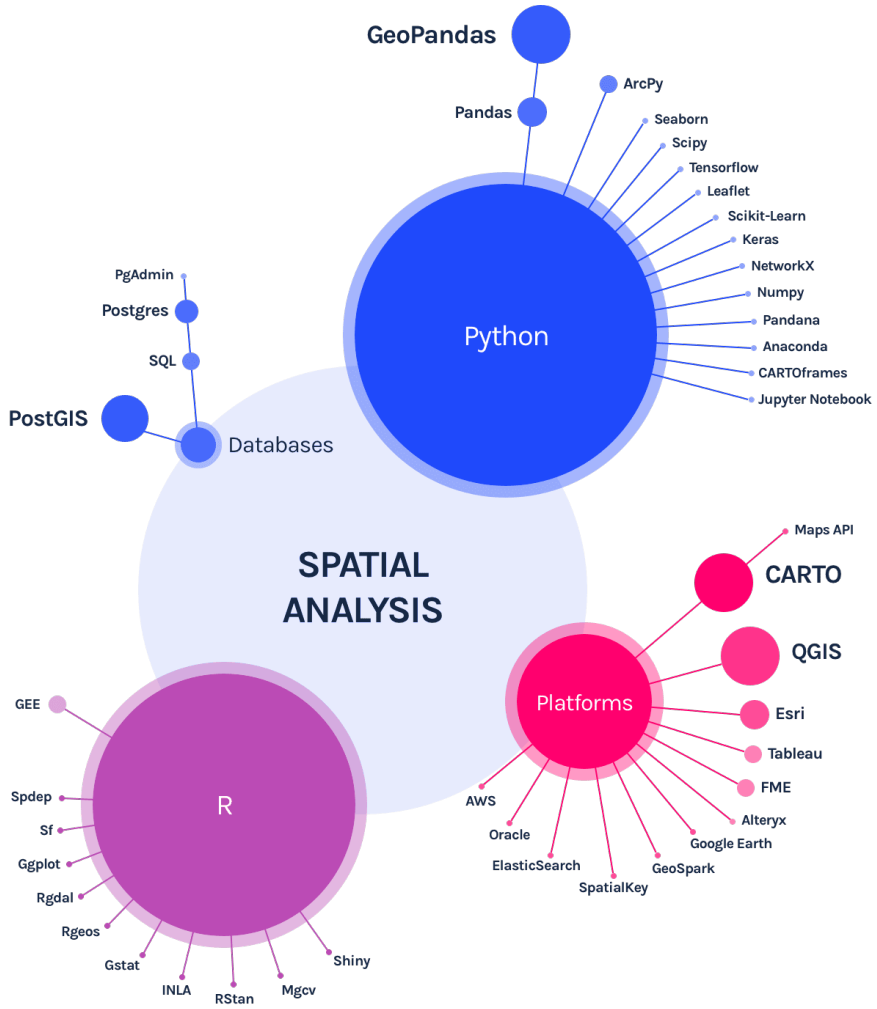
Technologie

polars

pandas

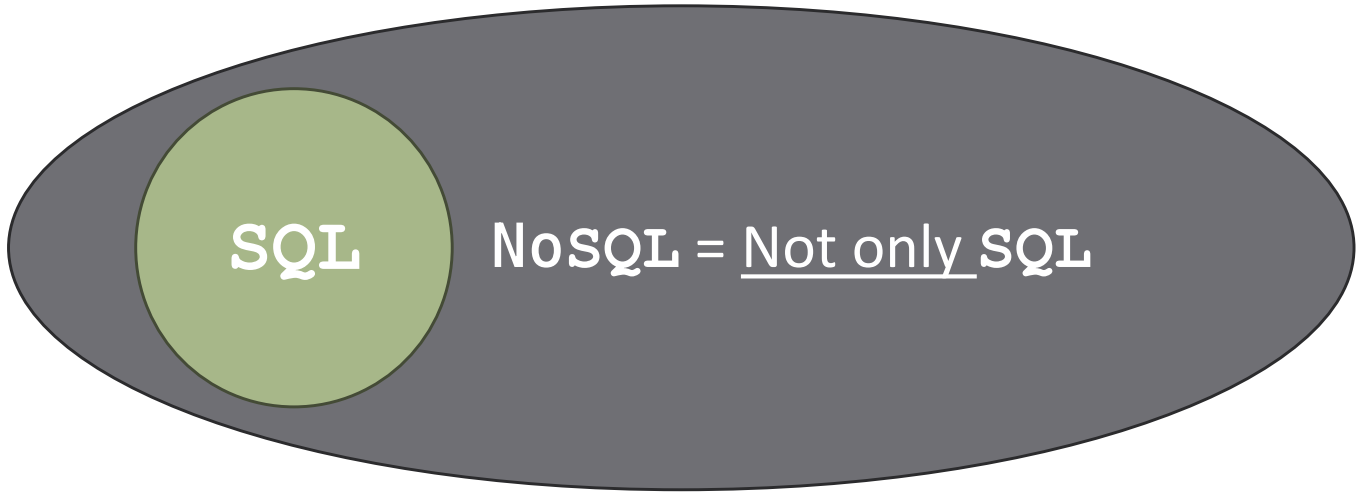
DASK
Distributed Summit 2021

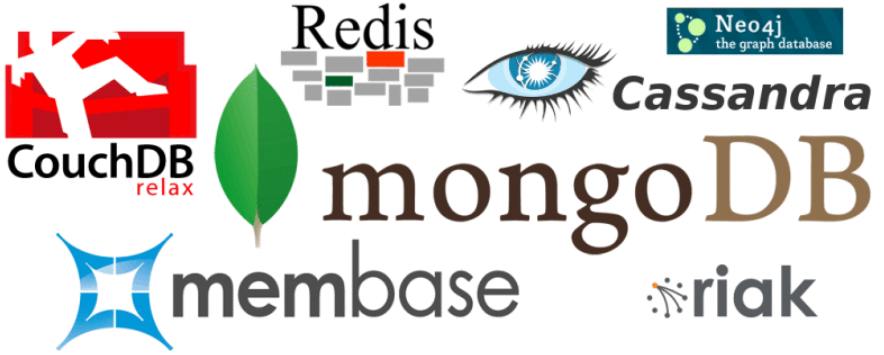
vaex 4.0.0-alpha.13
documentation



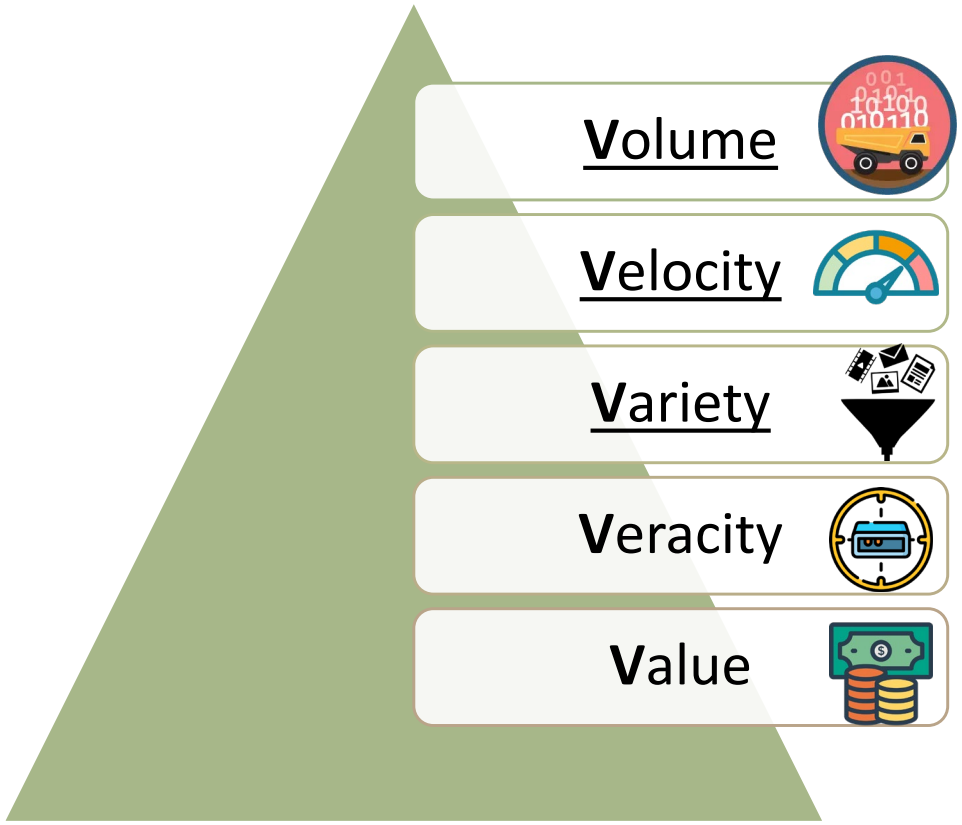
Duży bałagan (ang. *complexity*)

Technologie





Modele 3V i 5V



Czy jestem w stanie **zmieścić** dane na dostępnych nośnikach?

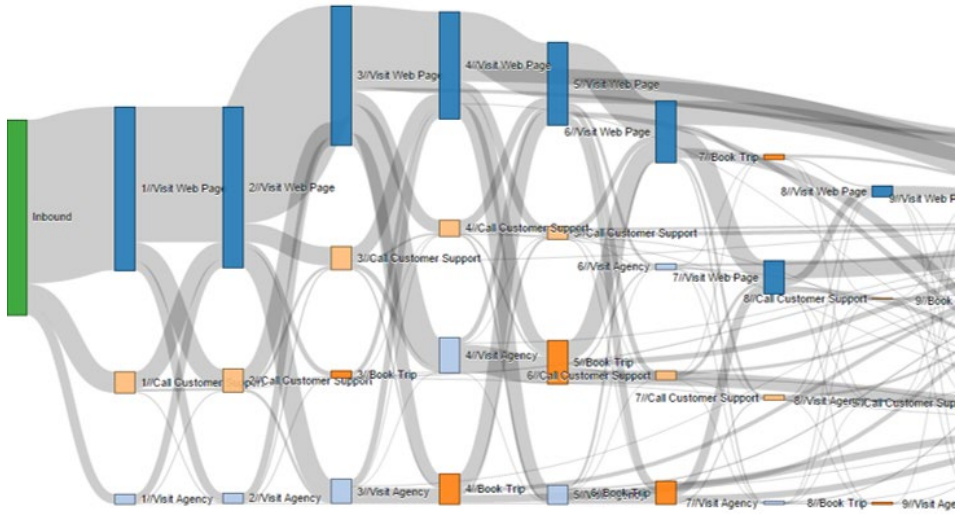
Jak **szybko** nadchodzą nowe dane?

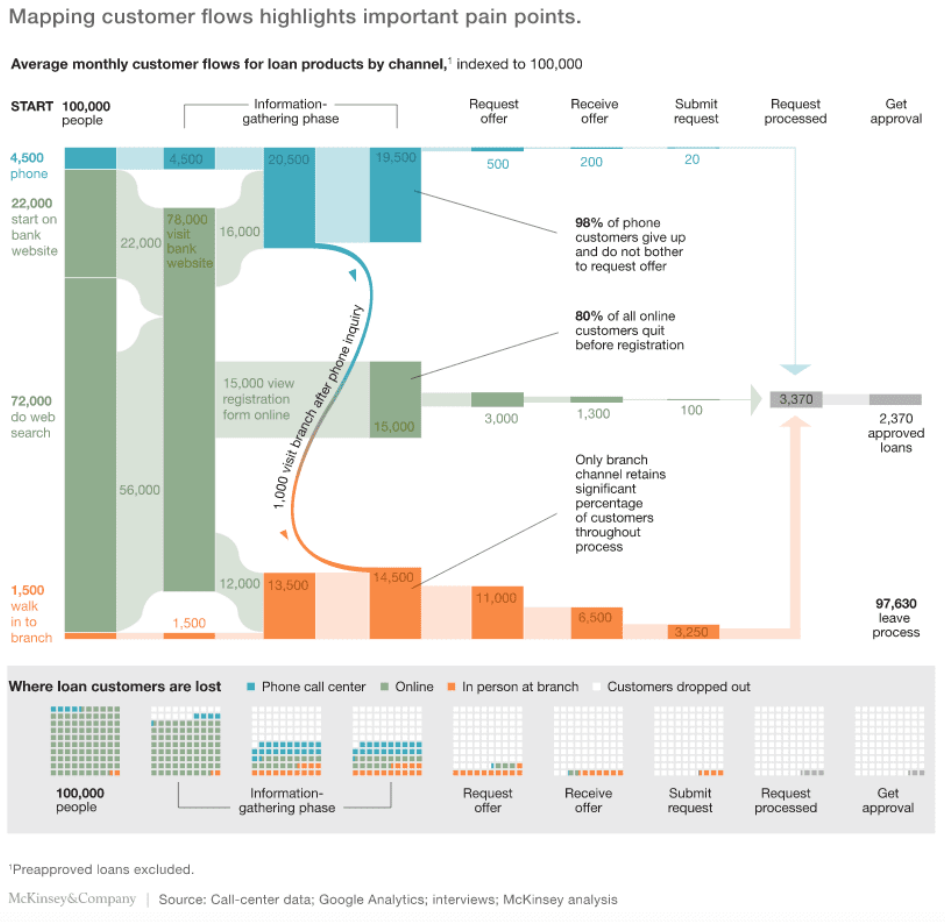
Jaki jest format / struktura / **źródła** danych?

Czy dane są **wiarygodne** / kompletne / spójne?

Czy **koszt** przetwarzania danych **się zwróci**?

Scieżki zakupowe

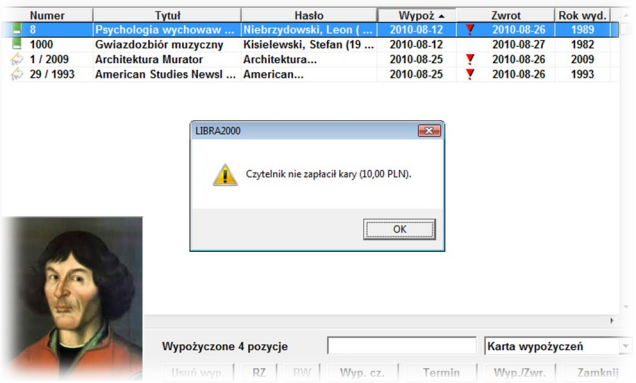




- Porzucone koszyki
- Zakupy krzyżowe, czego nie kupują
- Prognozowanie zamówień – stany magazynowe
- Segmentacja klientów, profilowanie (jakich filtrów używa)
- Odejście klienta – skąd pozyskany, niezbalansowane klasy
- Deduplikacja klientów (różne loginy *Neo4j*)
- Wielkość sprzedaży towaru – istotność towaru dla całości sprzedaży

Modele przetwarzania danych w ekosystemie Big Data

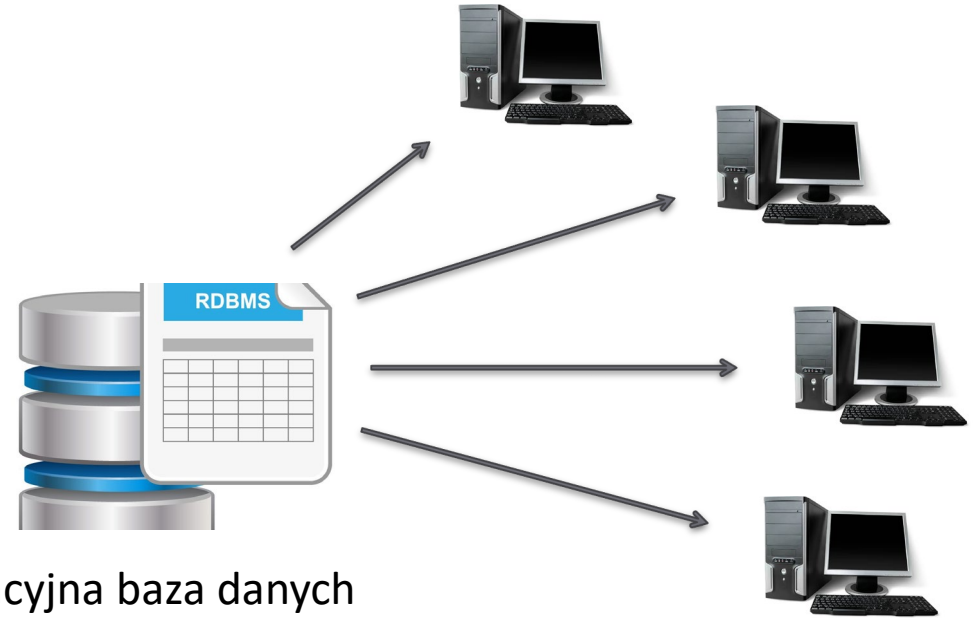
Kiedyś to było…



Relacyjna baza danych

Przetwarzanie informacji

Kiedyś to było…

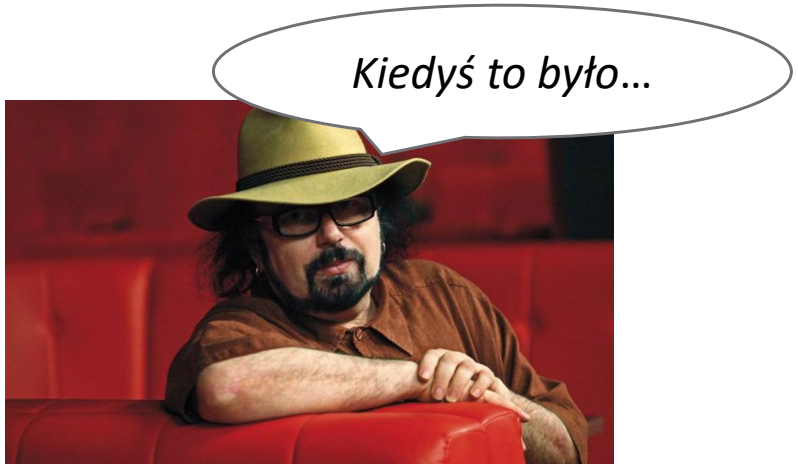


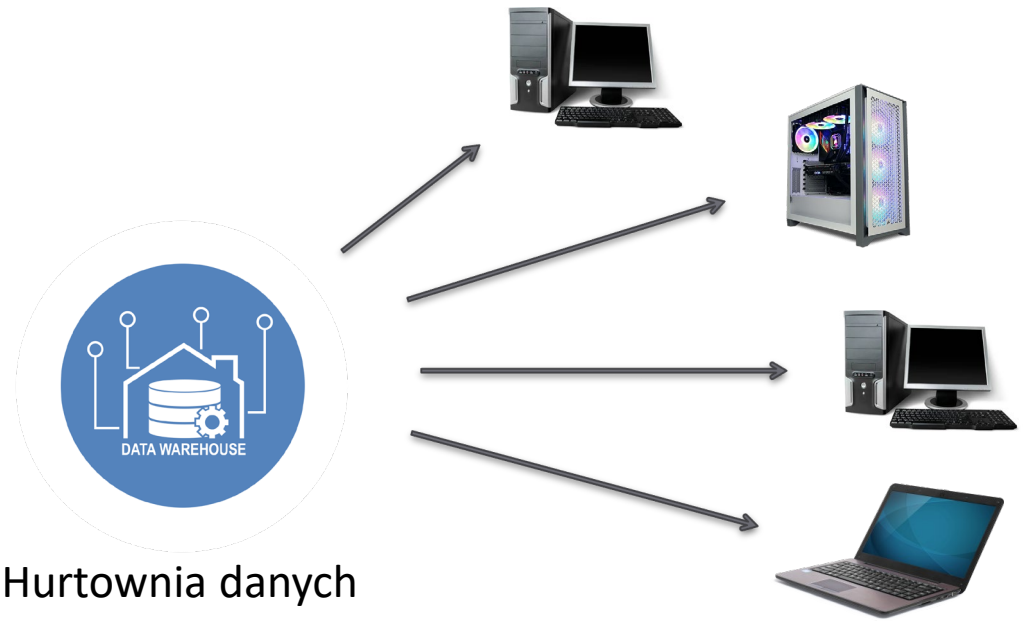
Relacyjna baza danych

• • •

NAS

***N**etwork **A**ttached **S**torage*





Hurtownia danych

• • •

Rozproszony system obliczeniowy

- Wiele komputerów widzianych przez użytkownika jako jedna całość

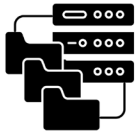
Skalowanie pionowe

W dłuższej perspektywie
trudno wdrożyć nowe technologie

Skalowanie p o z i o m e

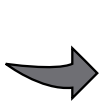
- wysokie koszty na wejściu w system zarządzania
- łatwo uzyskać niezawodność
- trudno spełnić warunki ACID dla transakcji
  ***A**tomicity*, ***C**onsistency*, ***I**solation*, ***D**urability*
- warunki BASE
  ***B**asically **A**vailable*, ***S**oft state*, ***E**ventual consistency*
- język dostępu do danych komplikuje się

Rozproszony system obliczeniowy

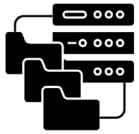


Rozproszony system plików

GFS



Rozproszony system obliczeniowy



Rozproszony system plików

System nierelacyjnych baz danych

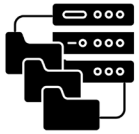
GFS

Cloud Bigtable

Amazon DynamoDB

mongo DB

Rozproszony system obliczeniowy



Rozproszony system plików

System nierelacyjnych baz danych

Środowisko przetwarzania rozproszonego

GFS → HDFS

Cloud Bigtable

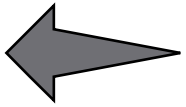
Amazon DynamoDB

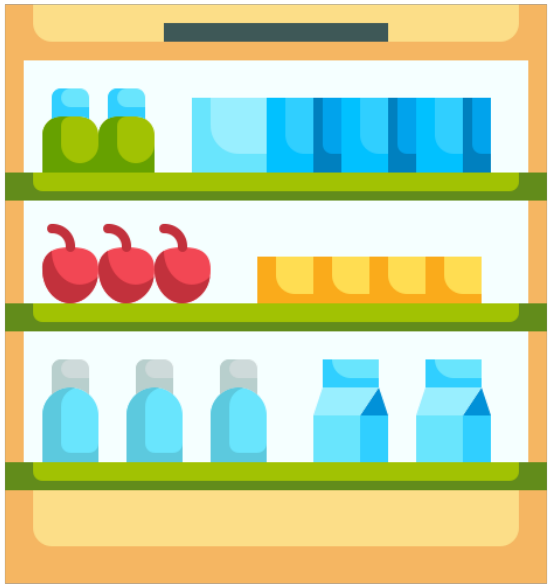
mongo DB

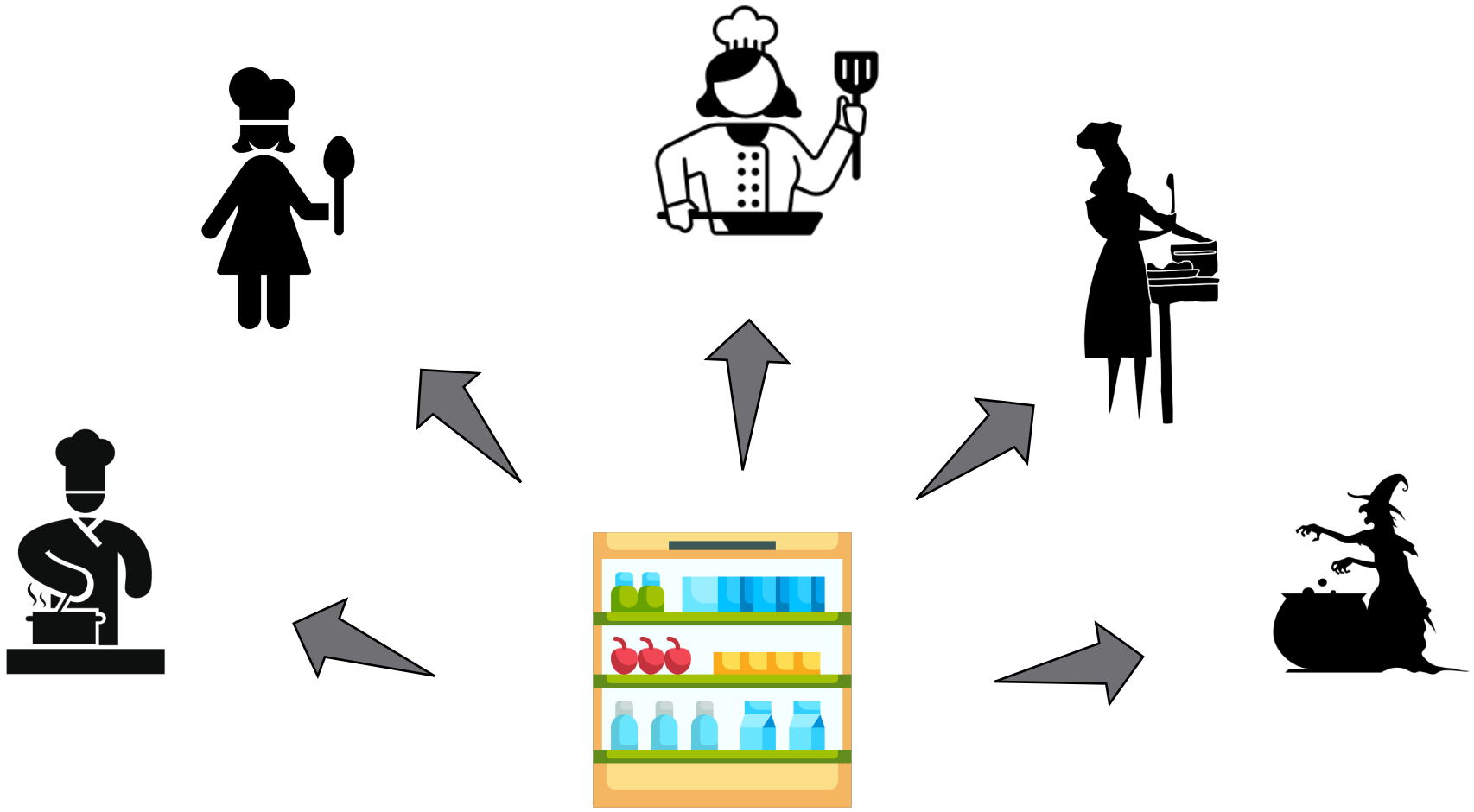
hadoop MapReduce

APACHE Spark™

modele programowania







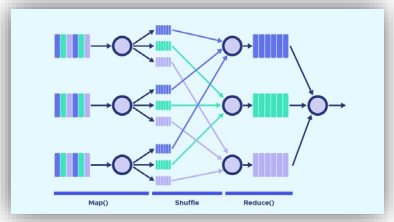
masterchef

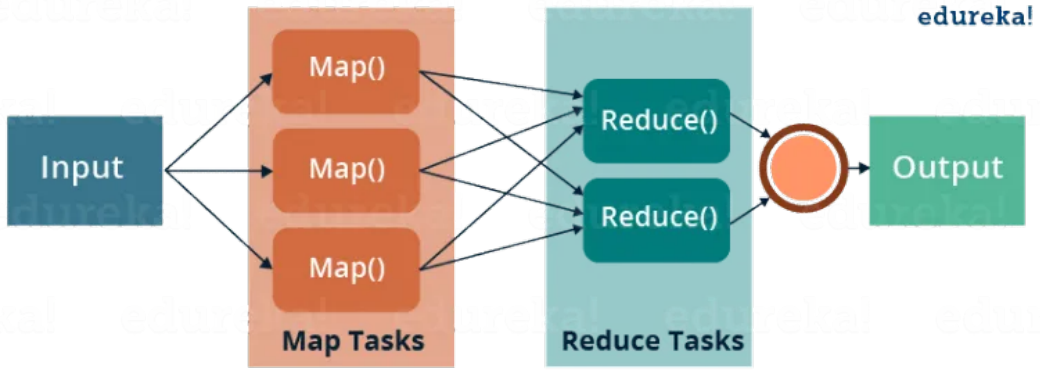
MASTER CHEF

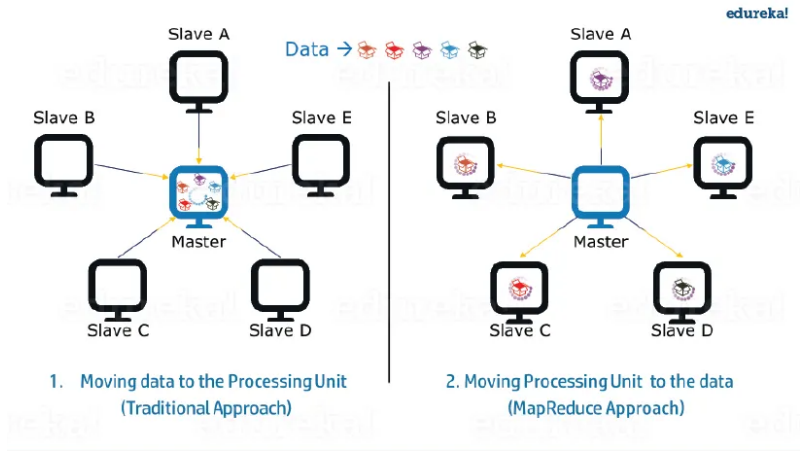
MAP

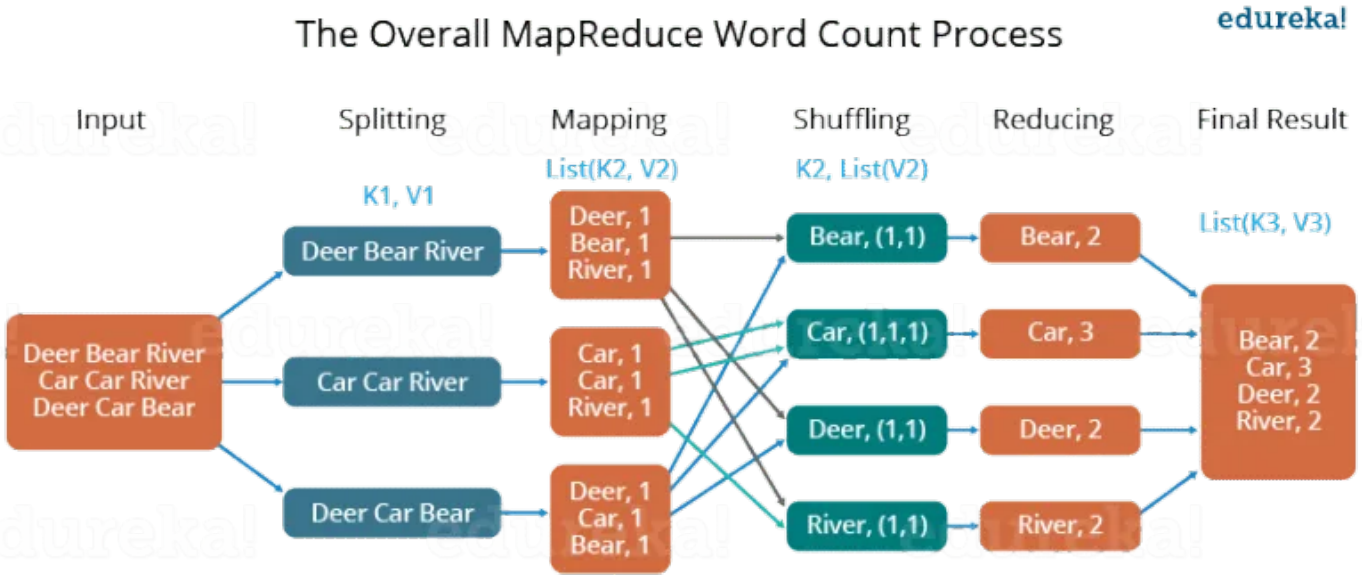
REDUCE

Model przetwarzania **MapReduce**



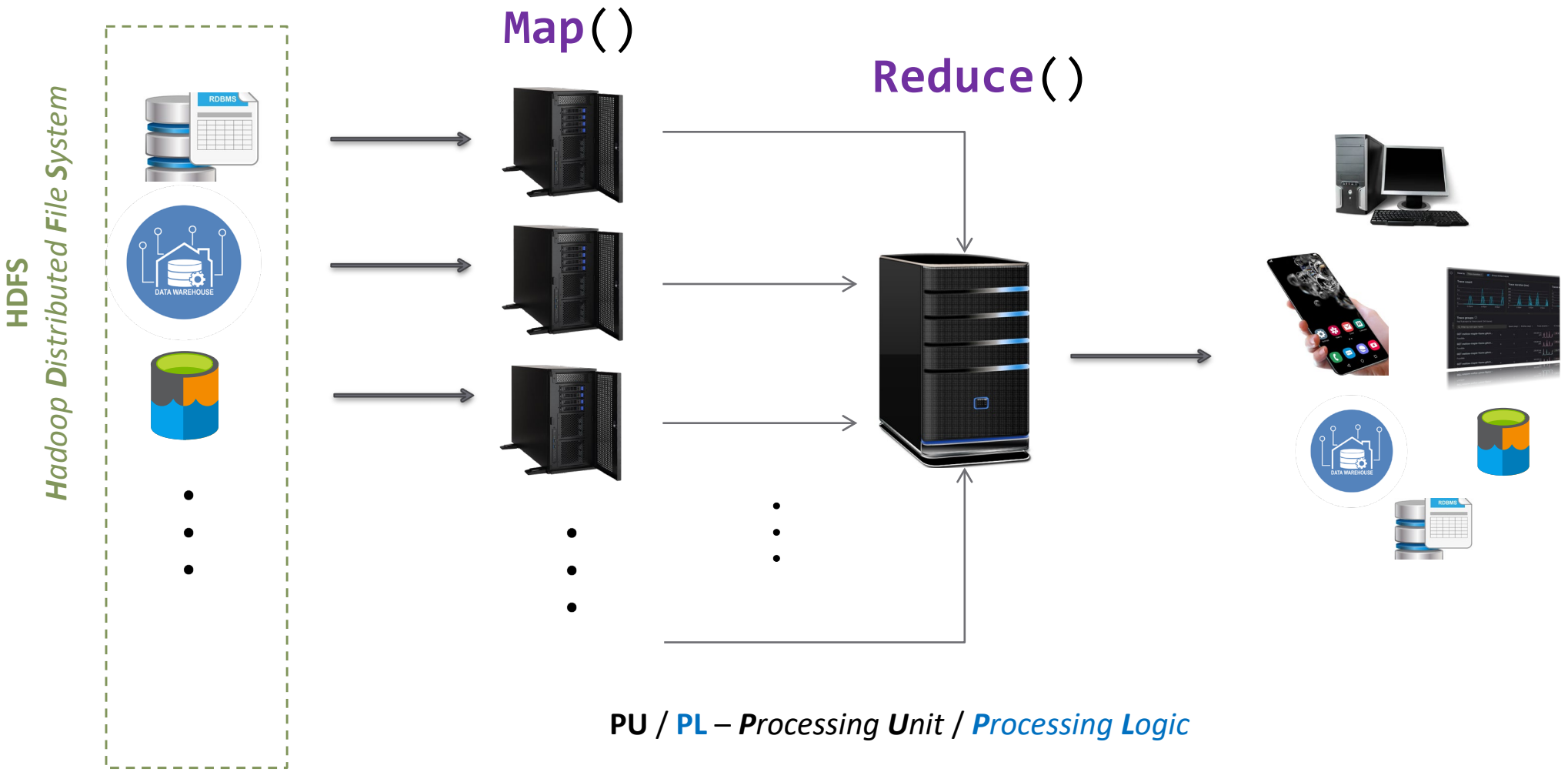




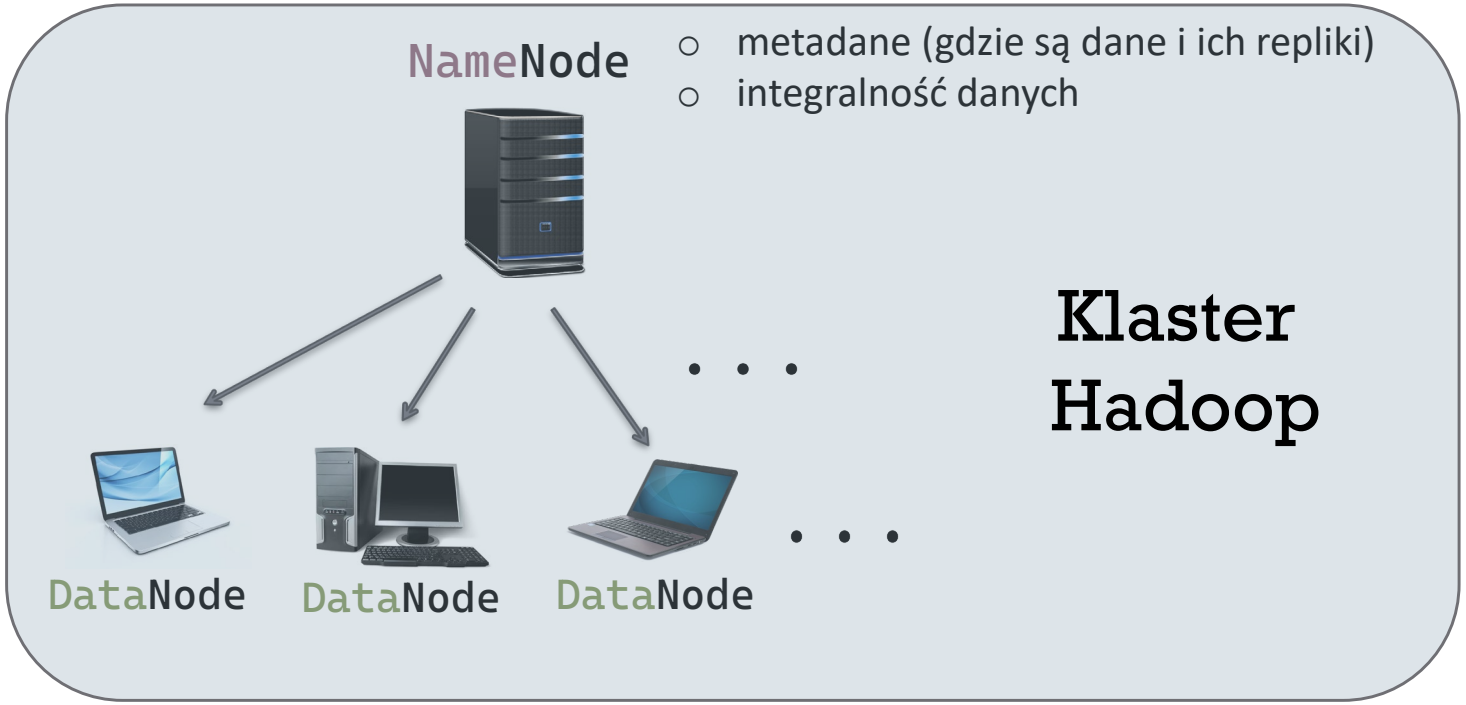


Dean, Jeffrey, and Sanjay Ghemawat. "*MapReduce: Simplified data processing on large clusters*." (2004).

Model **MapReduce**



PU / **PL** – ***P**rocessing **U**nit* / ***P**rocessing **L**ogic*

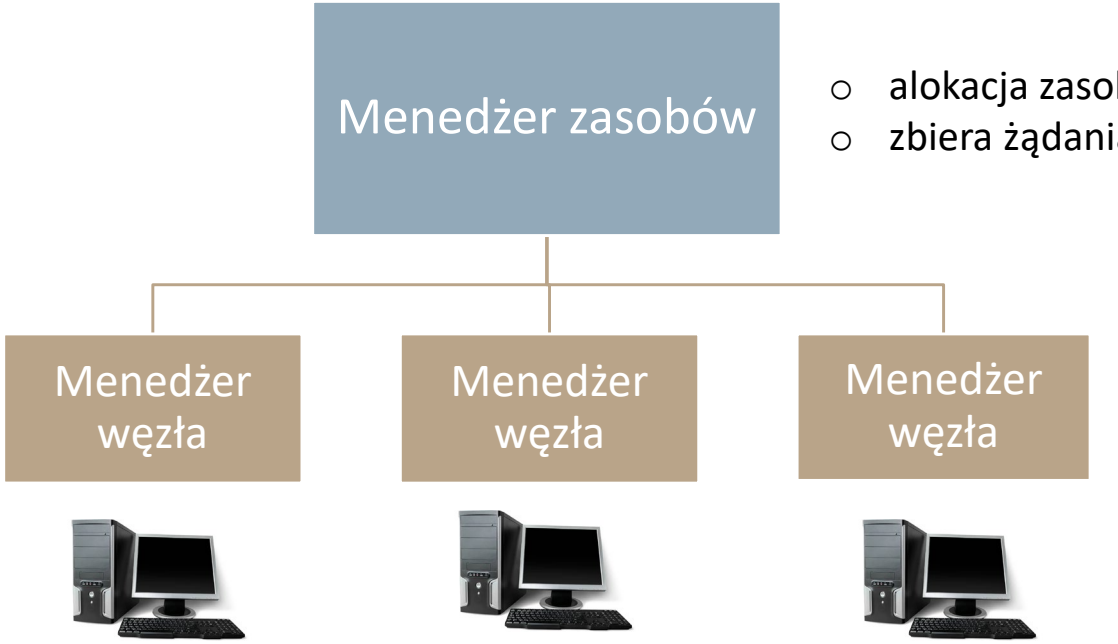


Rozwiązuje problemy:

- **V**olume
- **V**eracity + Variety

Yet Another Resource Negotiator

Organizuje przetworzenie ogromnej ilości danych w sposób efektywny



- alokacja zasobów do aplikacji
- zbiera żądania zadań i negocjuje pierwszy kontener

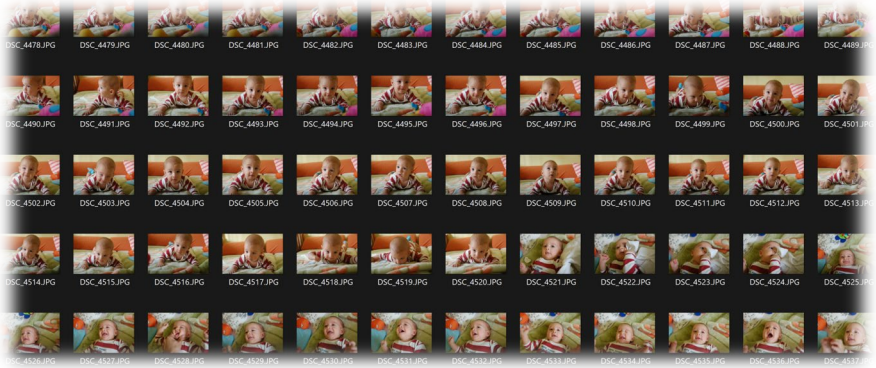
- Klient Hadoop (np. Java, Hive) tworzy **programy MapReduce**
- Programowanie funkcyjne

hadoop

Otwarta platforma do **przechowywania i przetwarzania** dużych wolumenów danych w **rozproszonych klastrach** bazujących na **ogólnodostępnym sprzęcie** komputerowym

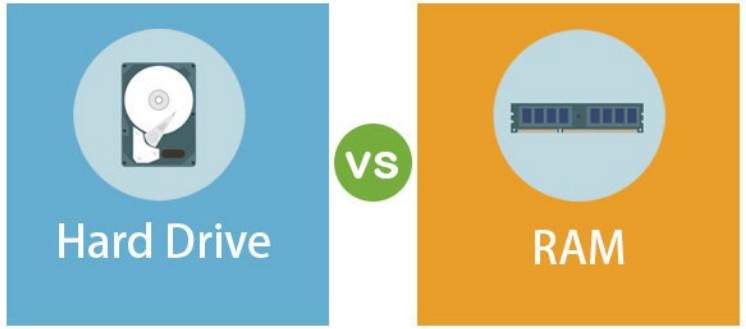
Podział plików + Podział zadań + Niezawodność

obsługa awarii sprzętu i danych



Koszt przechowywania bezużytecznych danych

Koszt filtrowania i usuwania



Model przetwarzania oparty na **danych** przechowywanych **na dyskach**

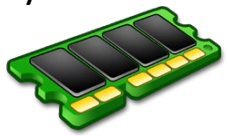
⇒

Nieefektywny w zadaniach **iteracyjnych** i **interakcyjnych**

Uczenie maszynowe

Przetwarzanie strumieni

Model przetwarzania oparty na **danych pośrednich** przechowywanych **w pamięci**



⇒

Spark MLlib

RDD

DataFrame

	Website	Visited
0	datagy.io	2023-01-23
1	google.com	2023-01-24
2	bing.com	2023-01-04

Otwarta platforma do **przetwarzania w czasie rzeczywistym** danych w rozproszonym systemie obliczeniowym

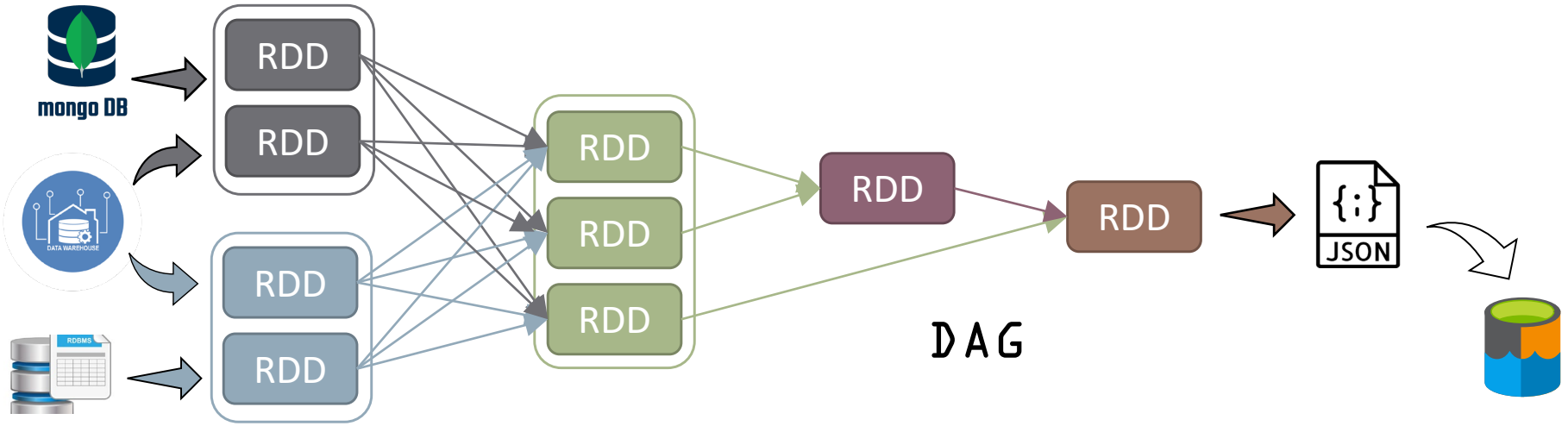
PySpark

Model przetwarzania w *Spark*

- Dane lub instrukcje ich wytworzenia są reprezentowane przez **RDD**

  ***R**esilient **D**istributed **D**ataset*

- RDD może być większy od dostępnej pamięci



- *DataFrame*
- *Dataset*

Podsumowanie Hadoop vs Spark

svitla

Key features	Apache Spark	Hadoop MapReduce
Speed	About 100 times faster than Hadoop	Faster than non-distributed systems
Data model	Use in-memory model to transfer data between RDD	Use HDFS to read, process and store large amounts of information (always use persistent storage)
Created in	Scala	Java
Style of processing	Batch, real-time, iterative, interactive, graph	Batch
Caching	Store data in memory	Caching data is not supported

AUDYCJA ZAWIERA LOKOWANIE PRODUKTU

Inne rozwiązania

databricks

Photon

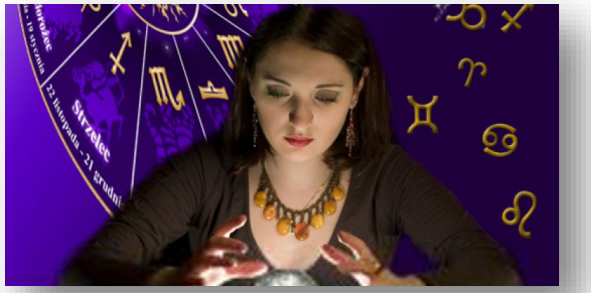
The next generation engine for the Lakehouse

Photon is the next generation engine on the Databricks Lakehouse Platform that provides extremely fast query performance at low cost – from data ingestion, ETL, streaming, data science and interactive queries – directly on your data lake. Photon is compatible with Apache Spark™ APIs, so getting started is as easy as turning it on – no code changes and no lock-in.

Massively scalable storage for demanding applications

Red Hat® Ceph® Storage, an open, massively scalable, simplified storage solution for modern data pipelines. Engineered for data analytics, AI/ML, and emerging workloads, Red Hat Ceph Storage delivers software-defined storage on your choice of industry-standard hardware.

Przyszłość



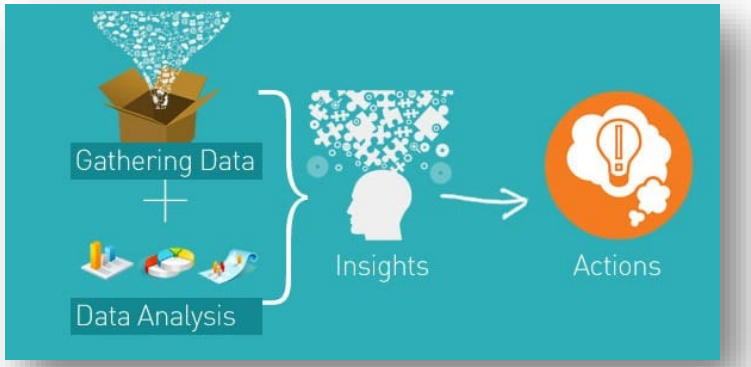
- Nowy model rozproszonej bazy danych (patrz *blockchain*)
  prawa własności ma twórca danych
- ***Interplanetary File System*** (IPFS) – sieć peer-to-peer z własnym DNS (IPNS) oraz wersjonowaniem plików

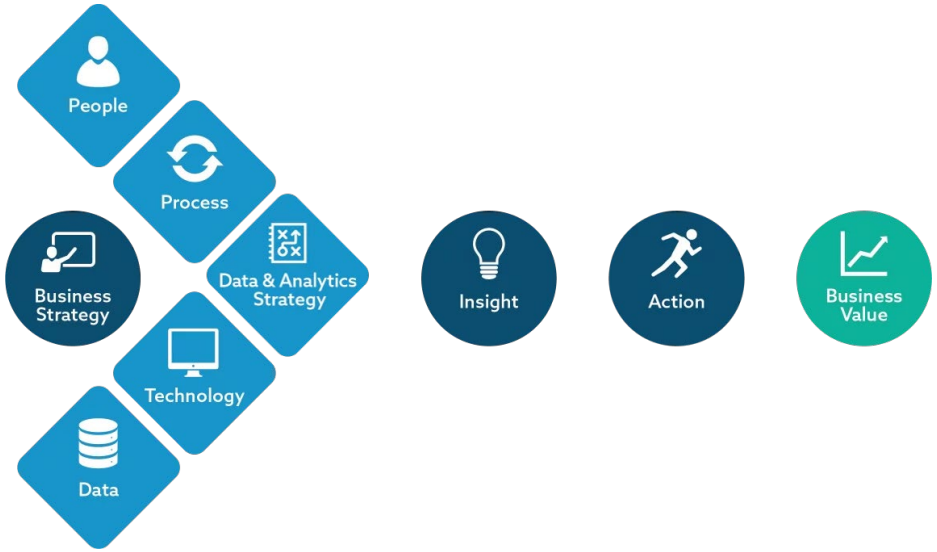
Filecoin

IPFS

Uwagi końcowe

Przetwarzanie danych w ekosystemie Big Data





Baza pozbawiona schematów = baza z elastyczniejszym schematem


THE
NEVERENDING
STORY