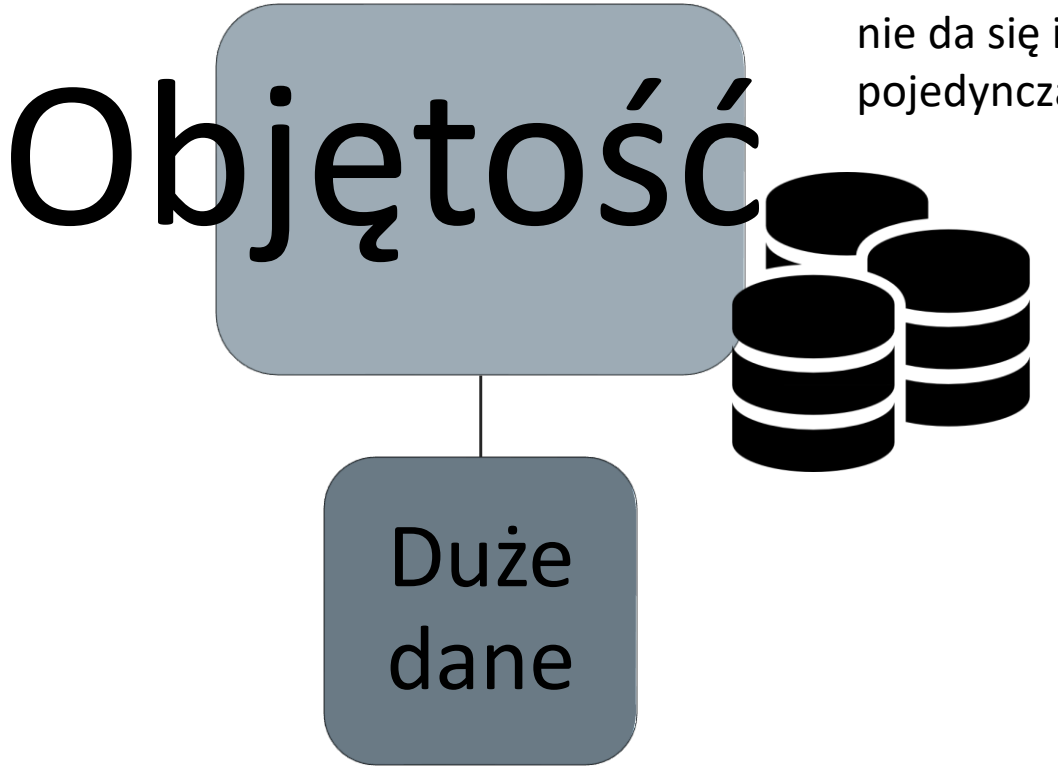
Metody i narzędzia *Big Data*

Modele przetwarzania danych

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**?

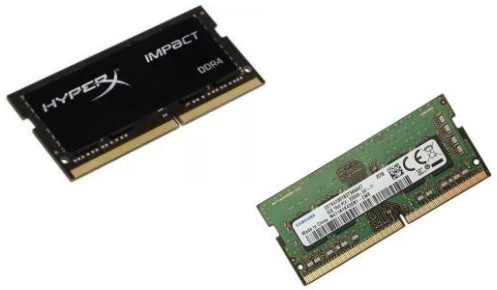


nie da się ich załadować na pojedynczą maszynę i przetwarzać

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**



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

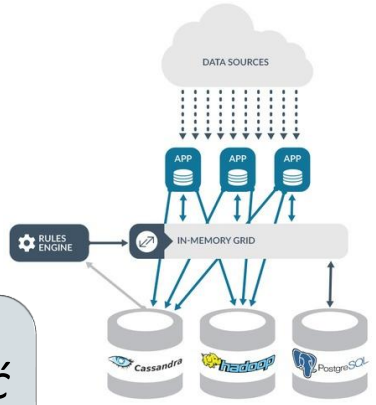
Co to znaczy, że dane są **duże**?

Objętość

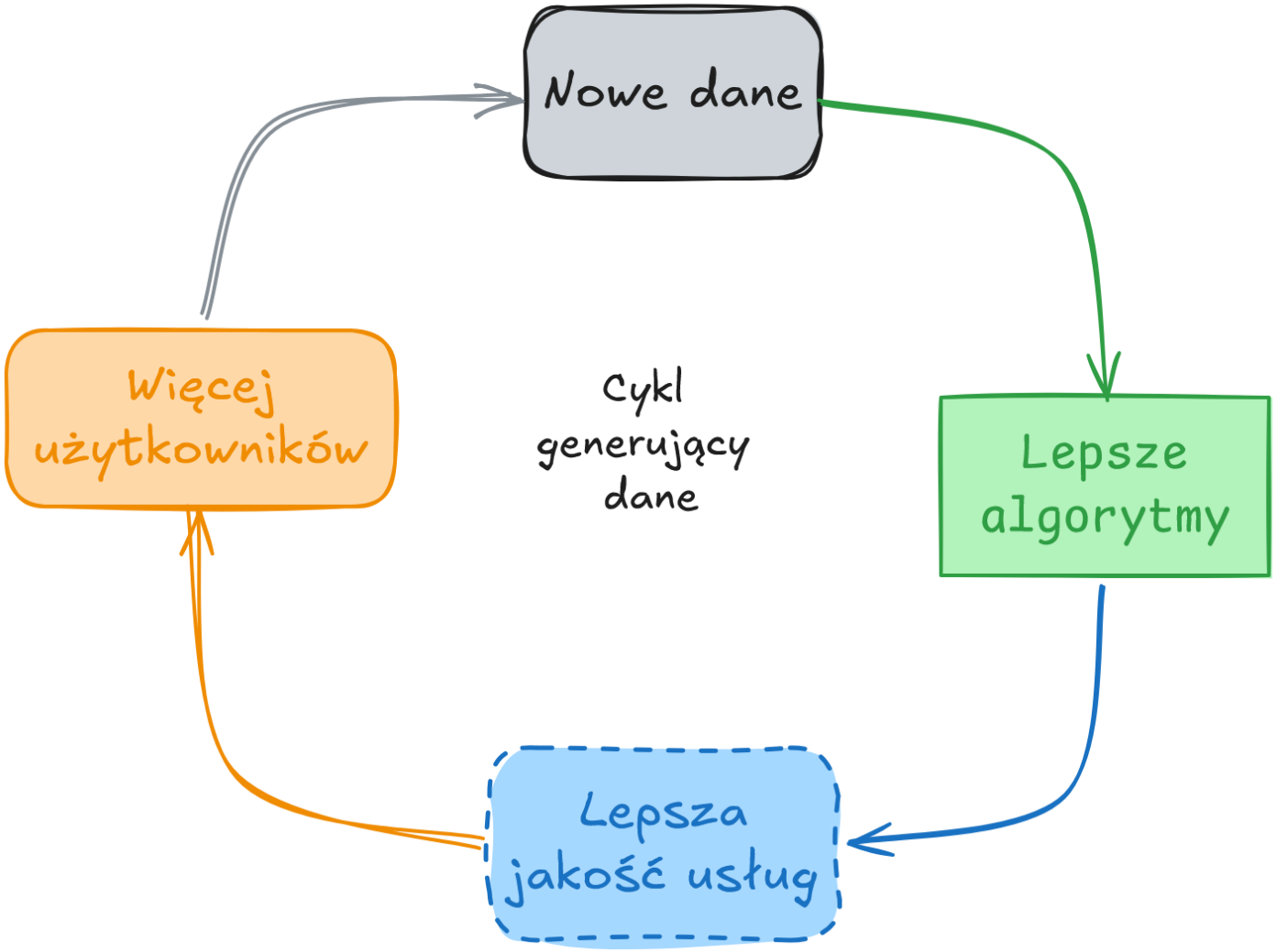
nie da się ich załadować na pojedynczą maszynę i przetwarzać

Duże dane

Szybkość i zmienność (strumienie)



Szybko narastają


Nowe dane
Lepsze algorytmy
Lepsza jakość usług
Więcej użytkowników
Cykl generujący dane

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

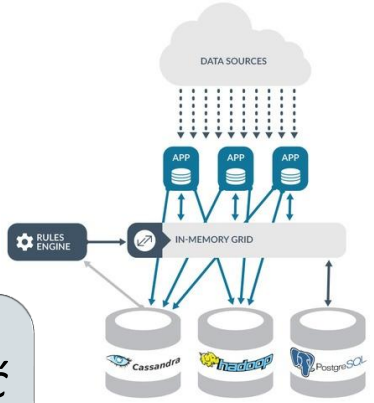
Co to znaczy, że dane są **duże**?

Objętość

Duże dane

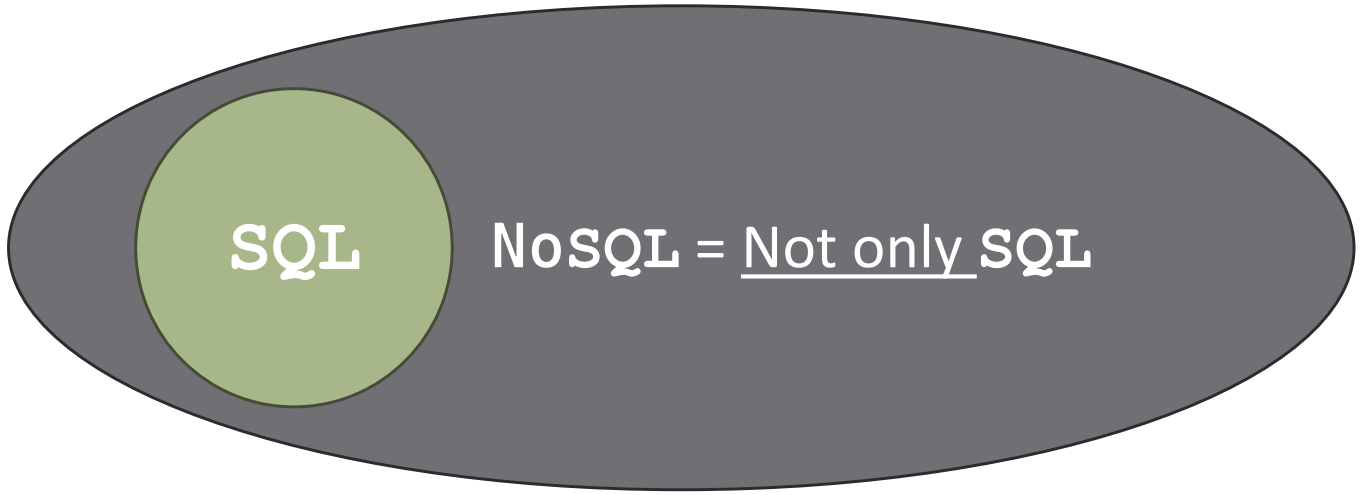
RÓnżoRodność i neipuorzdąkwoanie

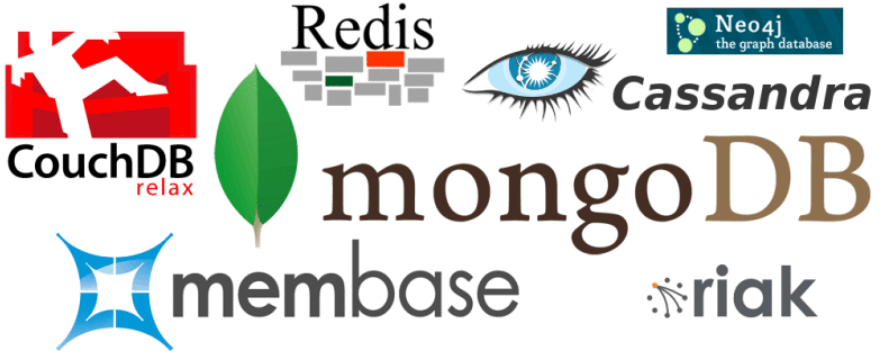
Szybkość i zmienność (strumienie)



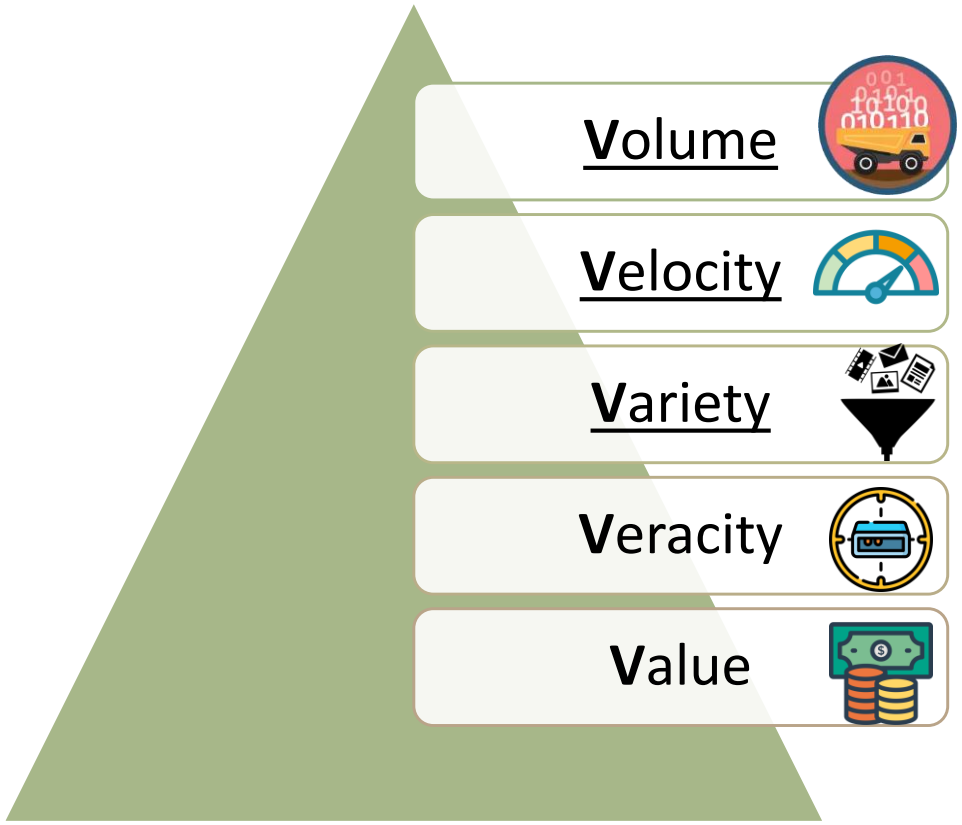
Szybko narastają

Duży bałagan (ang. *complexity*)





Modele 3V i 5V



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

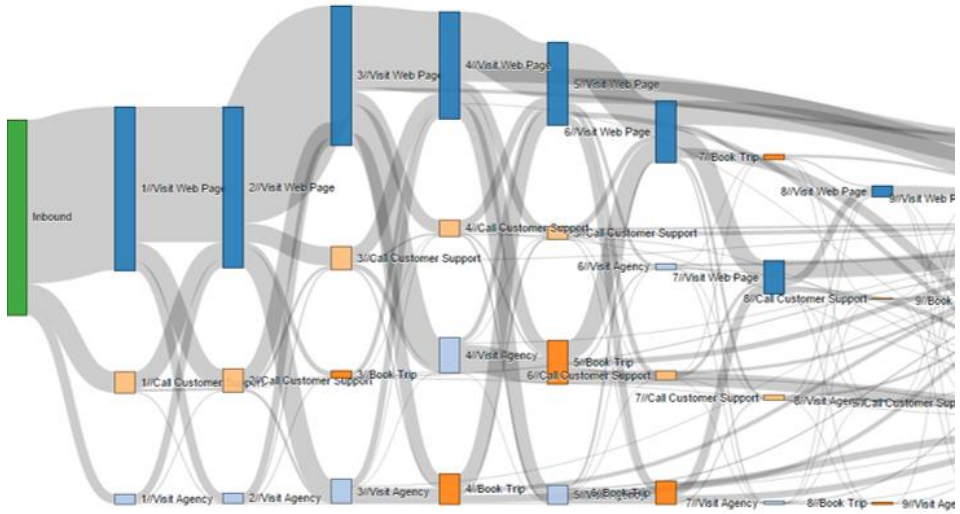
Jak **szybko** nadchodzą nowe dane?

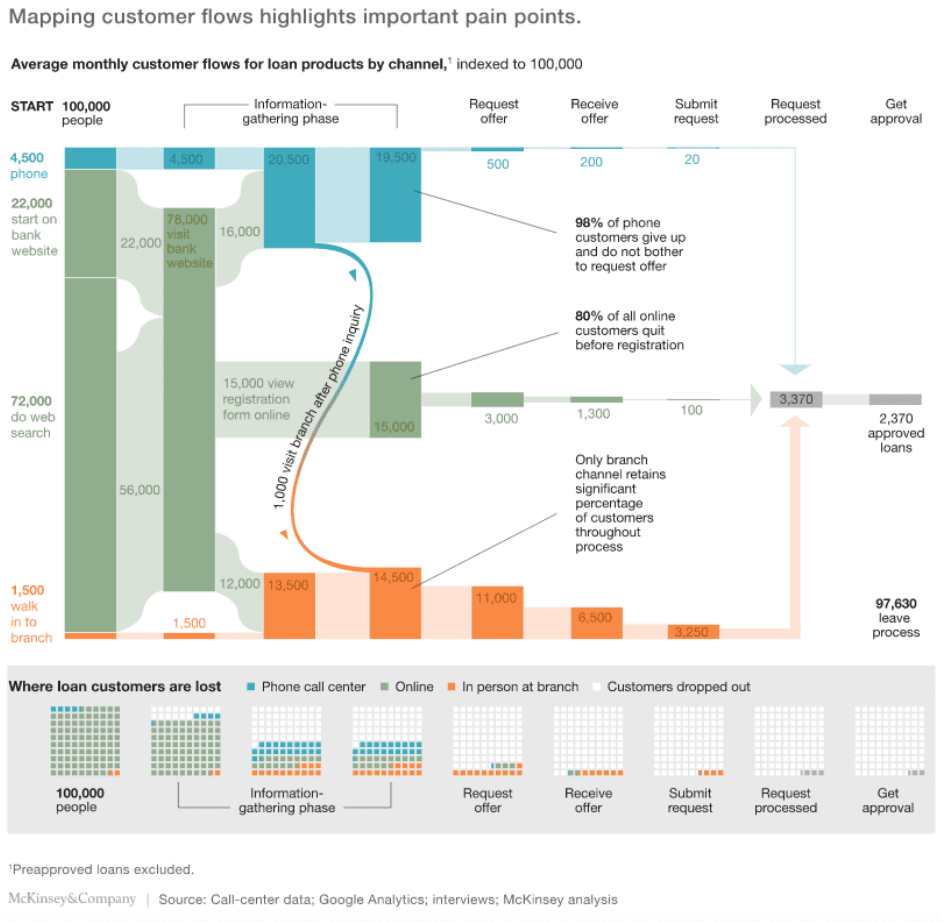
Jaki jest format / struktura / **źródła** danych?

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

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

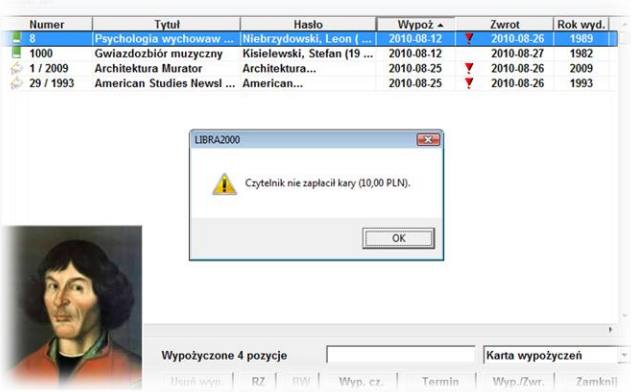
Scieżki zakupowe





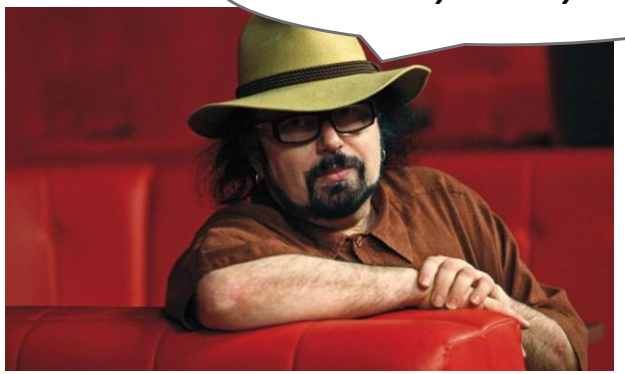
- Porzucone koszyki
- Zakupy krzyżowe, czego nie kupują
- Prognozowanie zamówień – stany magazynowe
- Segmentacja klientów, profilowanie (jakich filtrów używa)
- Odejście klienta – skąd pozyskany, niezbalansowane klasy
- Deduplikacja klientów (różne loginy *Neo4j*)
- Wielkość sprzedaży towaru – istotność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

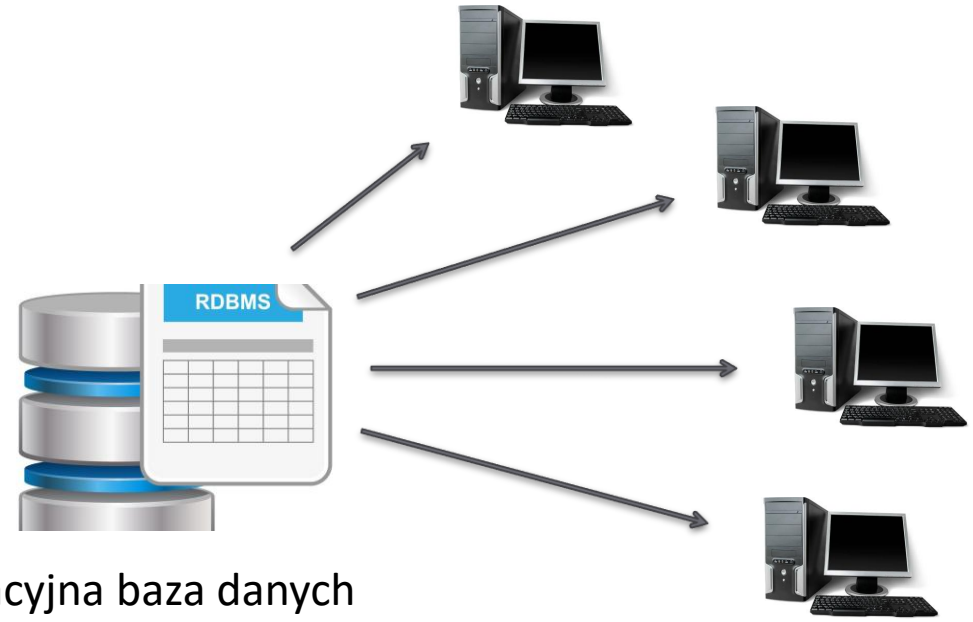
Kiedyś to było…



Relacyjna baza danych

Przetwarzanie informacji

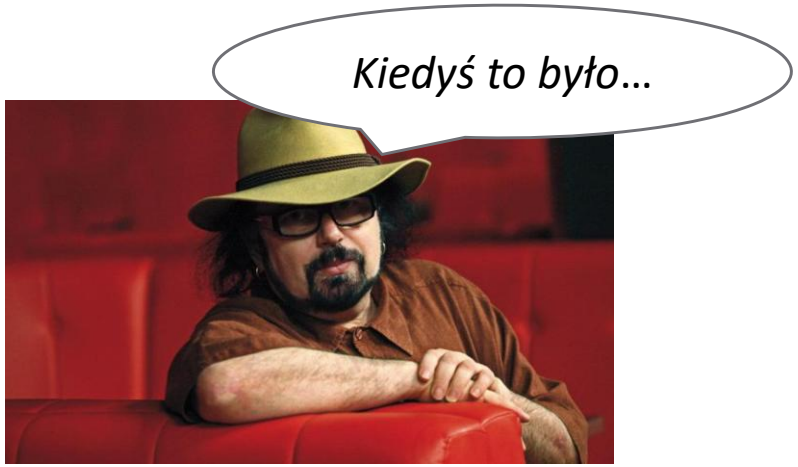


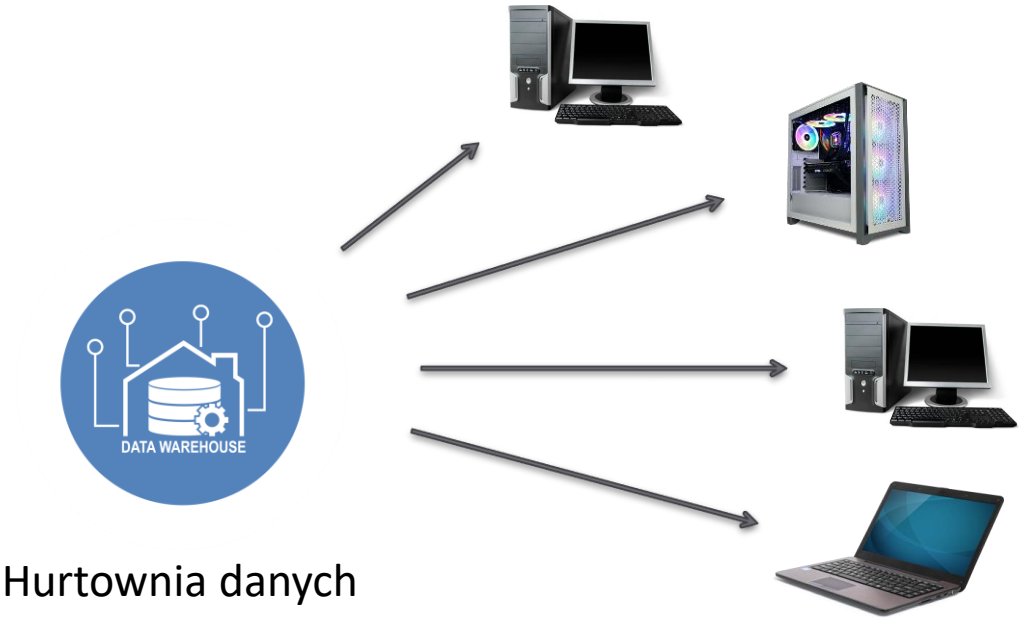


Relacyjna baza danych

• • •

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





Hurtownia danych

• • •

Rozproszony system obliczeniowy

- Wiele komputerów widzianych przez użytkownika jako jedna całość

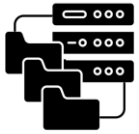
Skalowanie pionowe

W dłuższej perspektywie
trudno wdrożyć nowe technologie

Skalowanie p o z i o m e

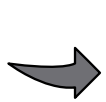
- Większa złożoność zarządzania
- Łatwiej uzyskać niezawodność przez replikację
- Trudno zapewnić wszystkie warunki ACID dla transakcji rozproszonych
  ***A**tomicity*, ***C**onsistency*, ***I**solation*, ***D**urability*
- Warunki BASE
  ***B**asically **A**vailable*, ***S**oft state*, ***E**ventually consistent*
- Dostęp do danych wymaga myślenia o partycjonowaniu, replikacji i opóźnieniach

Rozproszony system obliczeniowy

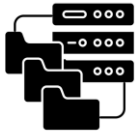


Rozproszony system plików

GFS



Rozproszony system obliczeniowy



Rozproszony system plików

GFS → HDFS → ceph

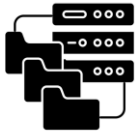
Nierelacyjne bazy danych

mongo DB

neo4j

redis

Rozproszony system obliczeniowy



Rozproszony system plików

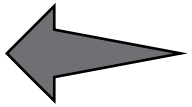
GFS → HDFS → ceph

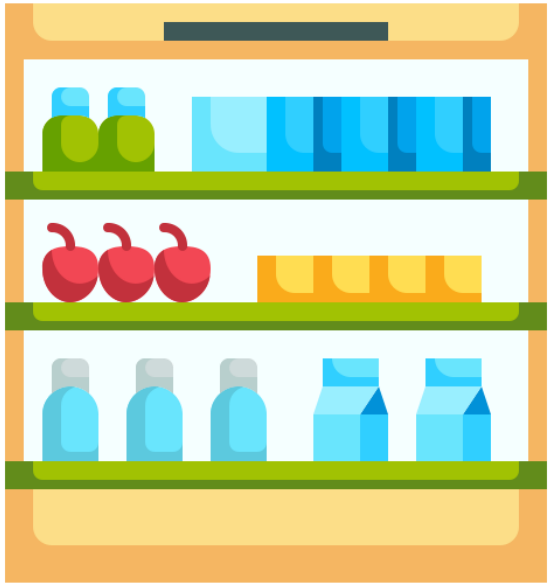
Nierelacyjne bazy danych

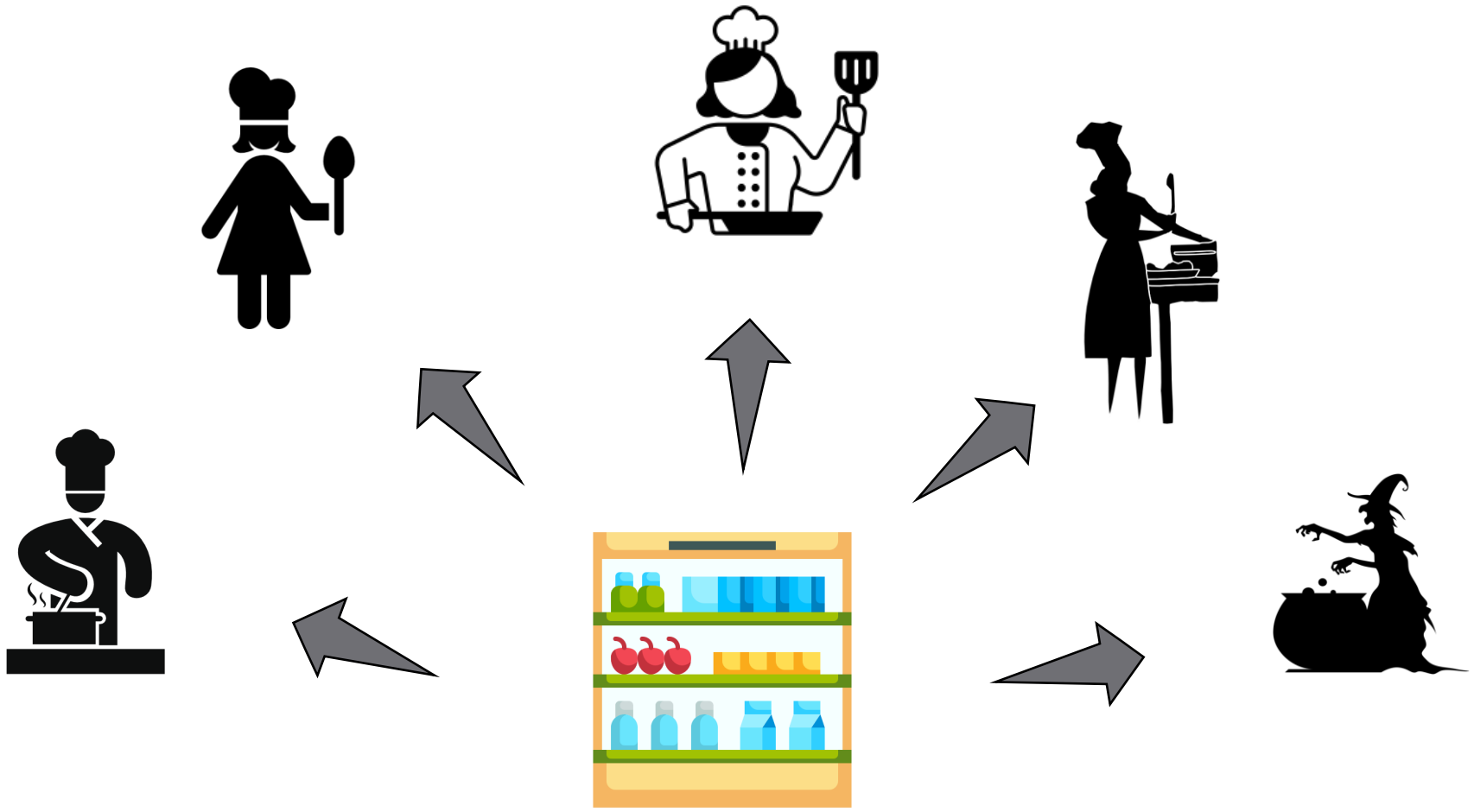
mongo DB, neo4j, redis

Platforma przetwarzania rozproszonego

hadoop Map Reduce, APACHE Spark





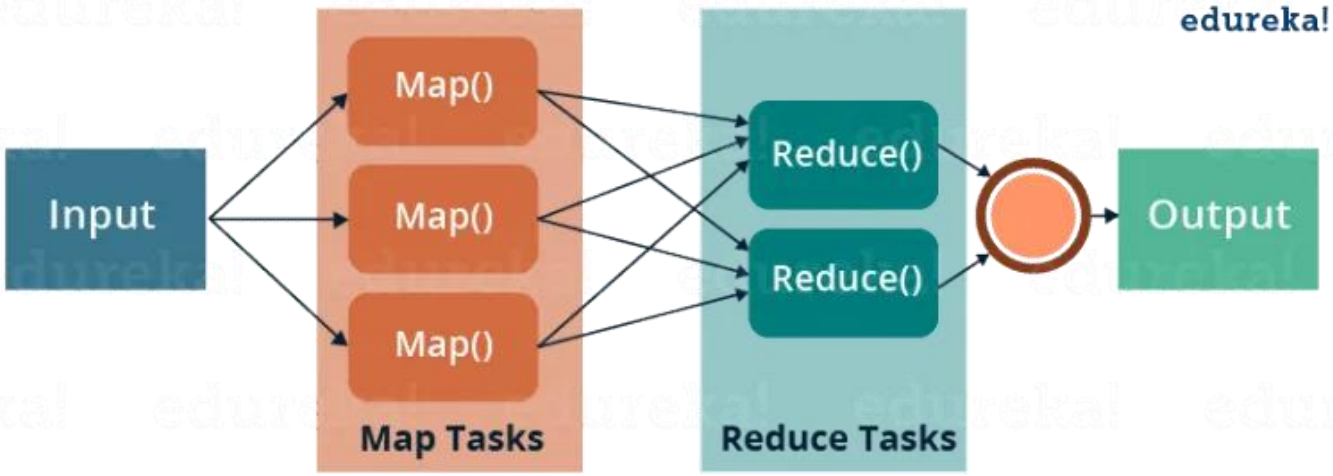


masterchef

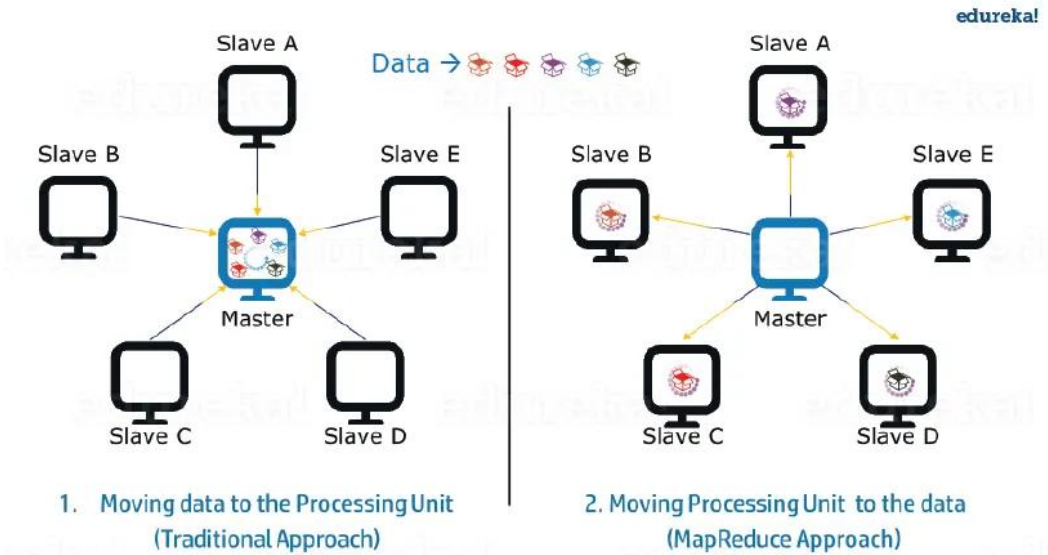
MAP

REDUCE

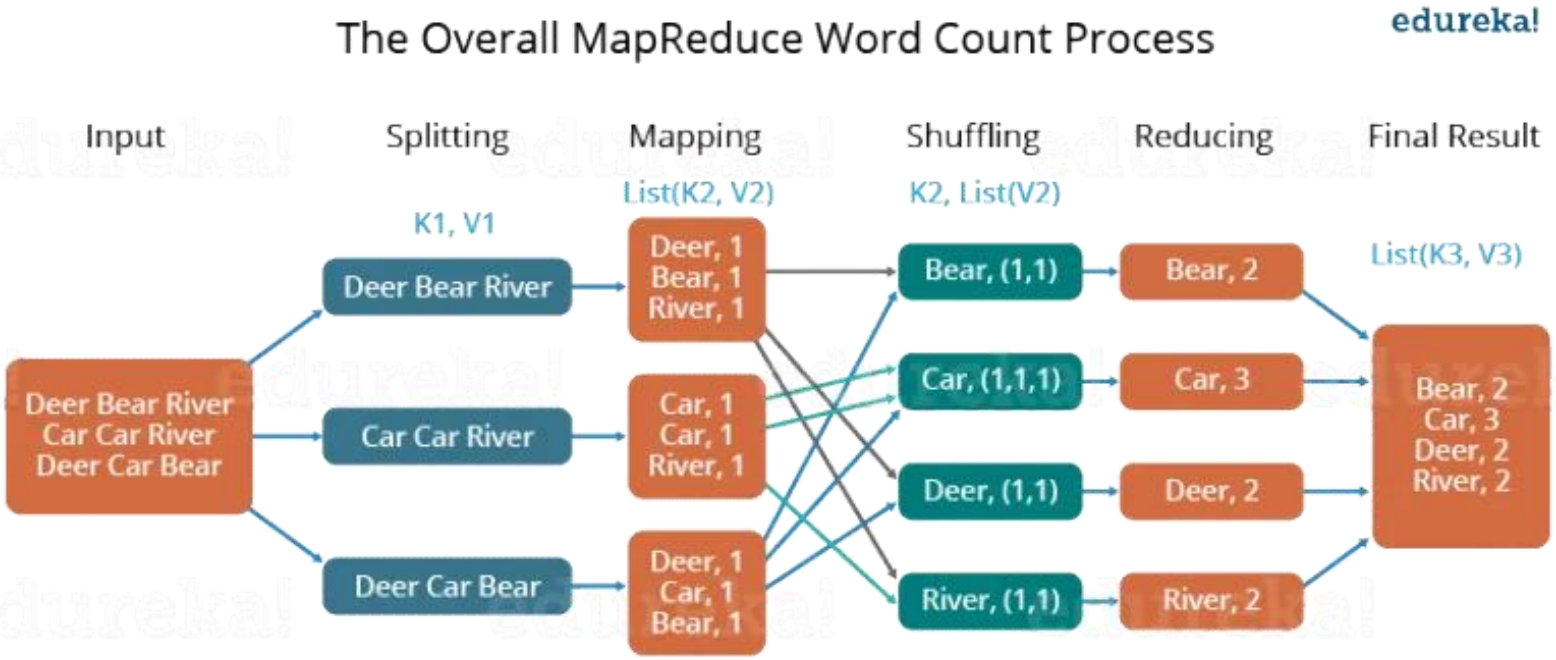
Model przetwarzania **MapReduce**



Model przetwarzania **MapReduce**

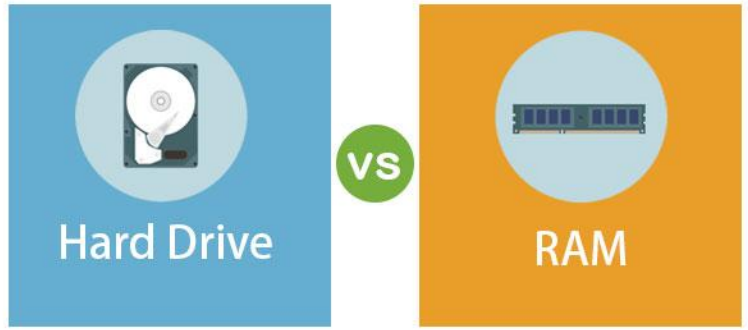


Model przetwarzania **MapReduce**



Koszt przechowywania bezużytecznych danych

Koszt filtrowania i usuwania



- Dane pośrednie przechowywane **na dyskach**
- Przetwarzanie wsadowe

⇒

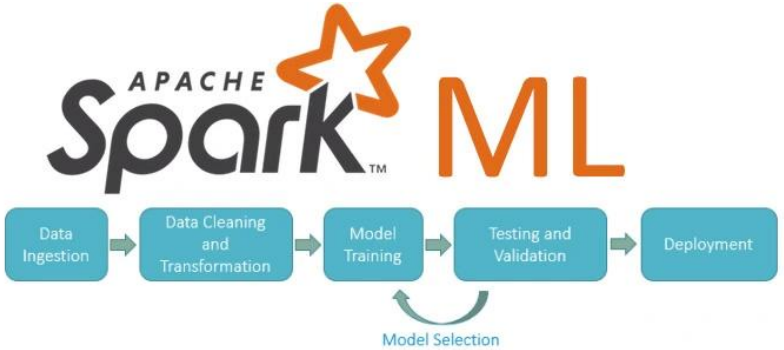
Nieefektywny w zadaniach **iteracyjnych** i **interakcyjnych**

Uczenie maszynowe

Przetwarzanie strumieni

- Dane pośrednie mogą być przechowywane **w pamięci**
- Szybsze iteracje

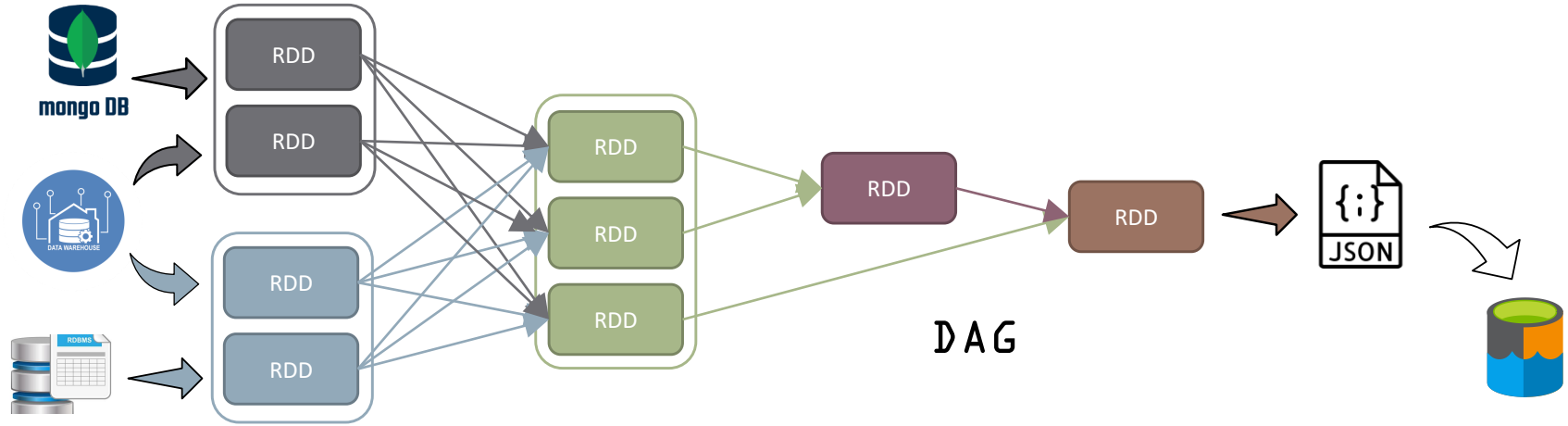
⇒



<u>Otwarta platforma</u> do szybkiego, rozproszonego przetwarzania danych, także **strumieniowych**

Model przetwarzania w *Spark*

- **RDD** (***R**esilient **D**istributed **D**ataset*) reprezentuje rozproszoną kolekcję danych oraz przepis na ich odtworzenie (ang. *lineage*)
- RDD może być większy od dostępnej pamięci
- Leniwa ewaluacja – DAG jest tworzony przy użyciu *transformacji*, obliczenia wykonuje się *akcjami*



Wybrane kierunki rozwoju

- Łańcuch bloków (ang. *blockchain)* – model rozproszonego rejestru danych ze współużytkowaniem, gdzie użytkownik może mieć większą kontrolę nad własnymi danymi
- ***Inter**planetary **F**ile **S**ystem* (IPFS) – rozproszony system przechowywania i **adresowania treści**: sieć peer-to-peer z systemem nazw IPNS oraz wersjonowaniem plików
- **Filecoin** – wynagradzanie za przechowywanie i udostępnianie przestrzeni dyskowej


THE END