Big Data

Organizacyjnie

Prowadzący:

dr Mariusz Rafało

mrafalo@sgh.waw.pl

http://mariuszrafalo.pl (hasło: BIG)

Zaliczenie:

Praca na zajęciach

Egzamin

Projekt

Plan zajęć

#	TEMATYKA ZAJĘĆ
1	Wprowadzenie do tematyki Big Data. Architektura Big data. Wybrane komponenty.
2	Przetwarzanie rozproszone: koncepcja, przykłady i zastosowania
3	Wybrane komponenty ekosystemu Big Data (technologie)
4	Analizowanie danych pozbawionych struktury
5	Analityka na platformie Big Data. Integracja ekosystemu Big Data z hurtownią danych.
6	Analizowanie danych w czasie rzeczywistym
7	Wybrane zagadnienia związane z etyką i prywatnością danych

Literatura

1. Provost, F. & Fawcett, T., **Data Science for Business: What you need to know about data mining and data-analytic thinking**, O'Reilly & Associates

2. Schutt, R. & O'Neil, C., **Doing data science**, O'Reilly

3. Minelli, M. **Big Data, big analytics: emerging business intelligence and analytic trends for today's businesses**, John Wiley & Sons, Inc.

4. Prajapati, V., **Big Data Analytics with R and Hadoop**, Packt Publishing Ltd.

5. Databricks, **A Gentle Introduction to Apache Spark**, Databricks

WPROWADZENIE

Definicja Big Data

- Big Data definiowane jest jako składowanie zbiorów danych o tak dużej złożoności i ilości danych, że jest to niemożliwe przy zastosowaniu podejścia tradycyjnego (np. opartego na hurtowni danych)
- Zagadnienie obejmuje identyfikację, pobieranie, składowanie, przeszukiwanie, współdzielenie, analizę i wizualizację danych

wikipedia.org

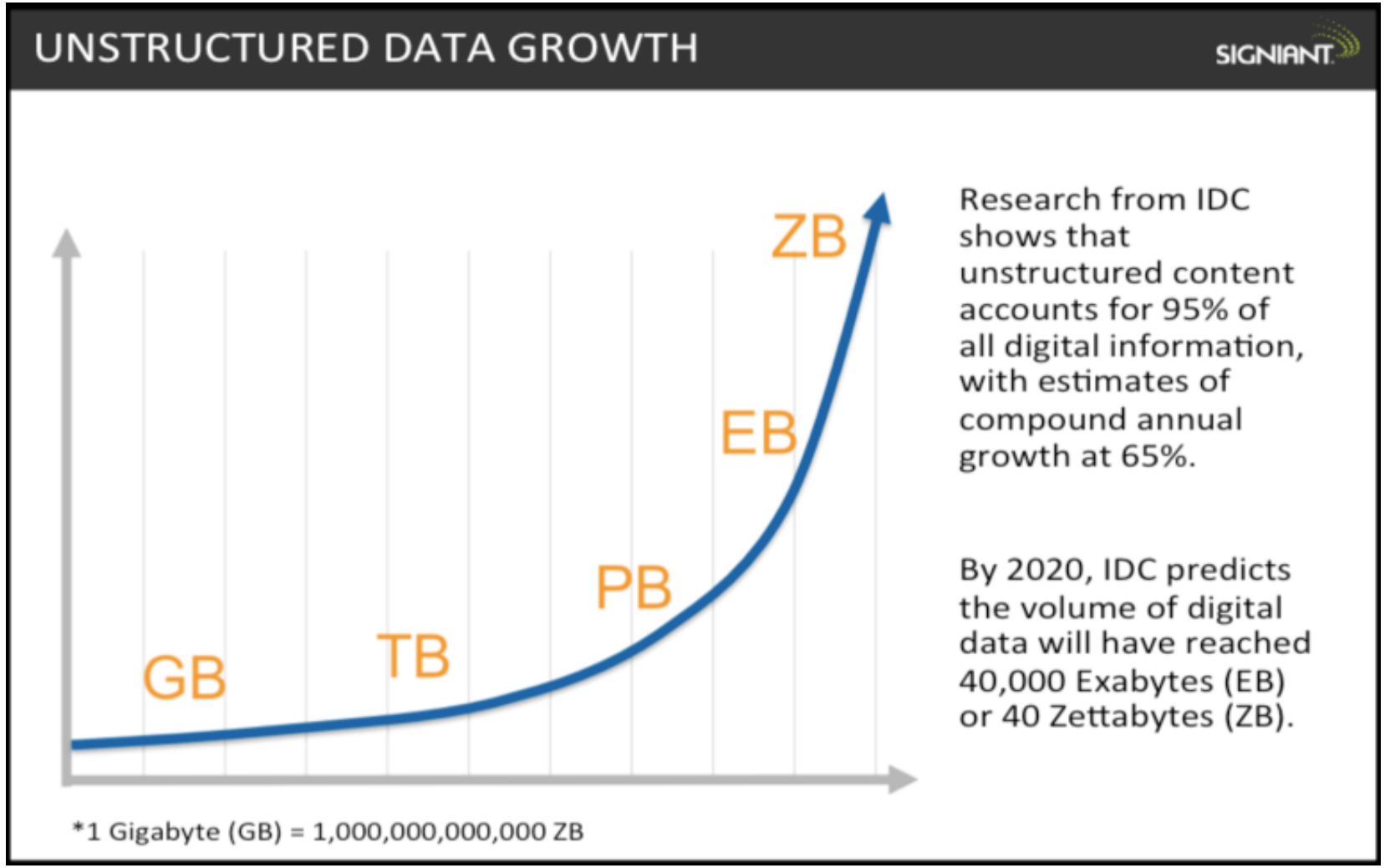
Paradygmat Big Data

Element	Paradygmat klasyczny	Paradygmat Big Data
Ilość danych	Kolejne przyrosty danych cyklicznie ładowane do hurtowni danych	Analizowanie danych w czasie rzeczywistym, zapisywanie wyłącznie informacji kluczowych
Szybkość danych	Cykliczne pobieranie wyłącznie istotnych danych. Wysoki (względnie) poziom latencji (opóźnienia) danych	Nasłuch strumienia danych, w momencie pojawienia się określonych sytuacji, natychmiastowe podjęcie działania
Różnorodność danych	Umieszczanie danych w bazie danych o określonej strukturze	Strukturyzowanie danych, które pozwalają na określenie kontekstu danych o nieustrukturyzowanej postaci

Big Data: 3V’s

- Wolumen danych (***Volume***)
- Zróżnicowanie danych (***Variety***)
- Szybkość zmian danych (***Velocity***)

Wolumen danych (Volume)

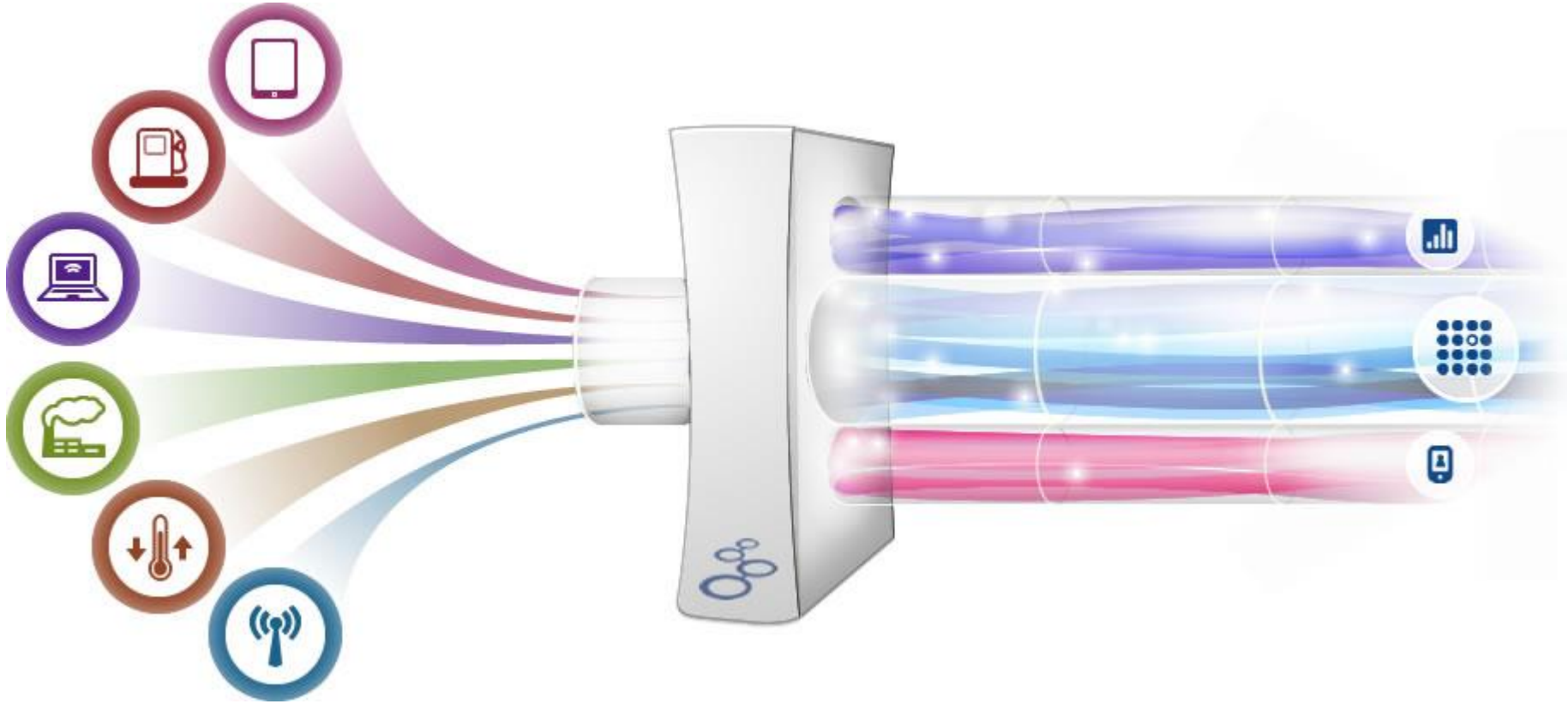


http://www.signiant.com

Zróżnicowanie danych (Variety)

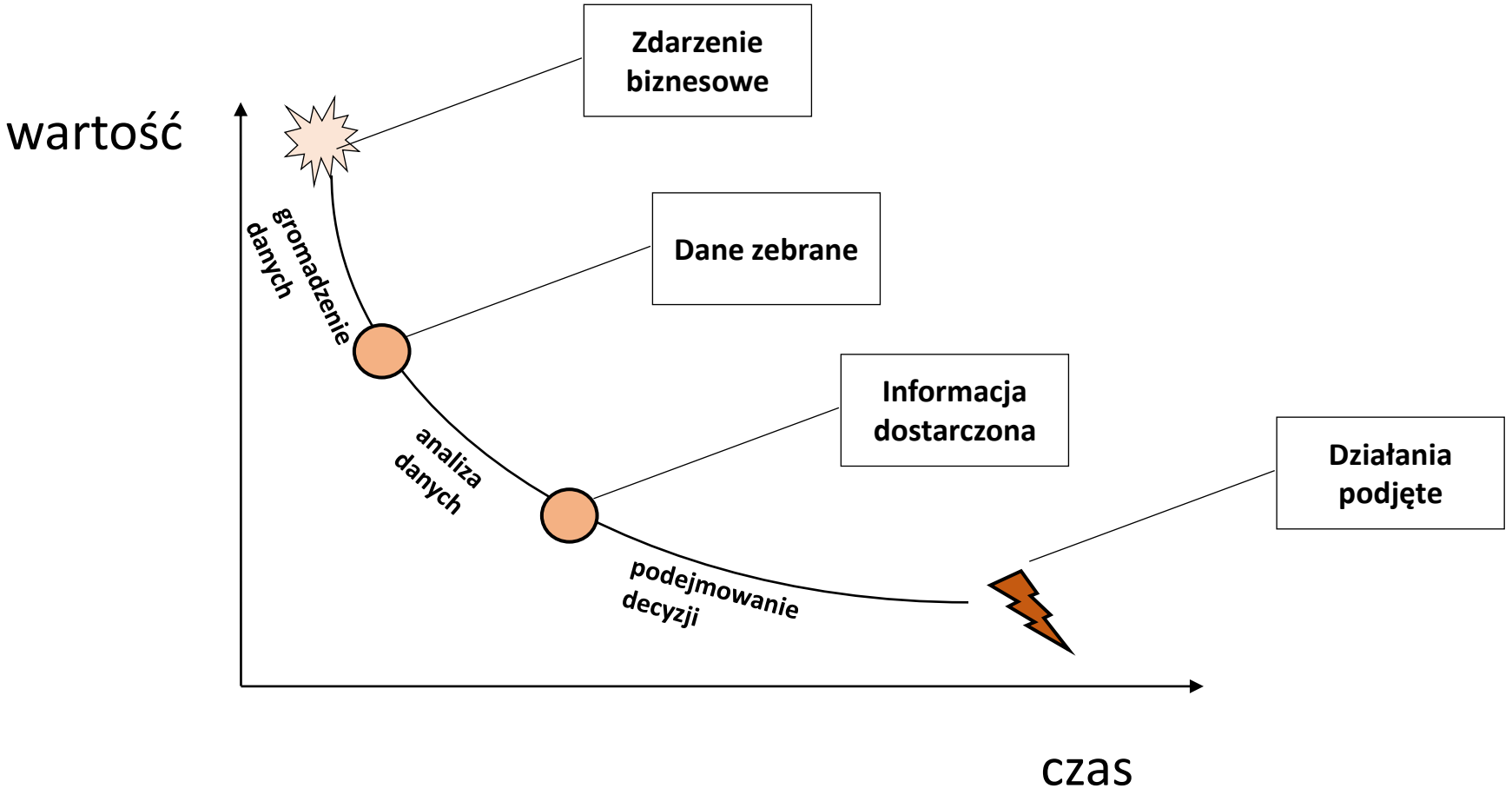
- Relacyjne bazy danych
- Tekst, html
- XML
- Strumienie danych
- Dane dotyczące powiązań
- Zdjęcia, filmy, muzyka

Szybkość danych (Velocity)



http://vceestartups.com

Big Data – kolejne V: Value (wartość)



Big Data – kolejne V: Veracity (wiarygodność)

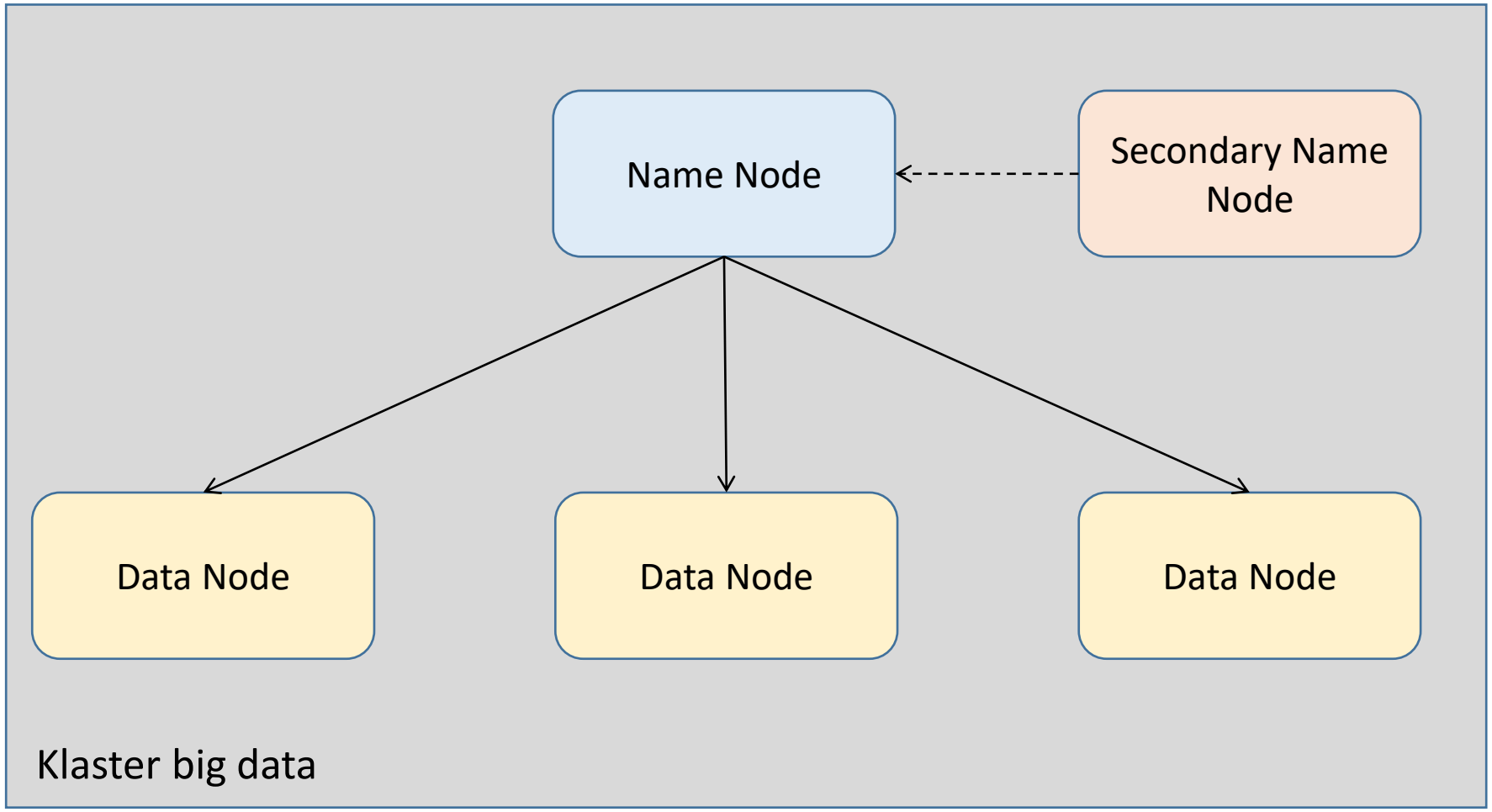
- Big data, ze względu na rodzaj danych oraz ich skalę, obarczony jest szeregiem problemów:
  - Błędy danych
  - Przekłamania
  - Szum informacyjny
  - Anomalie w danych
- W takich uwarunkowaniach istotne jest zarządzanie wiarygodnością danych dla ich użytkowników

Architektura Big Data

HDFS

- HDFS (*Hadoop Distributed File System*) to rozproszony system plików, umieszczony na wielu serwerach (węzłach – *node*)
- HDFS cechuje się wysokim poziomem tolerancji na awarie sprzętowe (*fault tolerant*)
- HDFS opiera się na nisko-kosztowych serwerach
- HDFS powstał na potrzeby projektu wyszukiwarki Nutch, dla firmy Yahoo

Architektura rozproszona



- Rozproszony system plików
- Automatyczny rebalancing
- Możliwość usuwania/wyłączania węzłów w trakcie pracy
- Brak systemu zabezpieczeń

Name node

- Przechowuje metadane plików; także dane dotyczące lokalizacji poszczególnych plików składowanych na HDFS
- *name node* jest kluczowym elementem architektury fizycznej – w klastrze zawsze jest jeden *name node*
- Zarządzania rozkładem plików podczas przyłączania nowych *data node* oraz w przypadku wystąpienia awarii

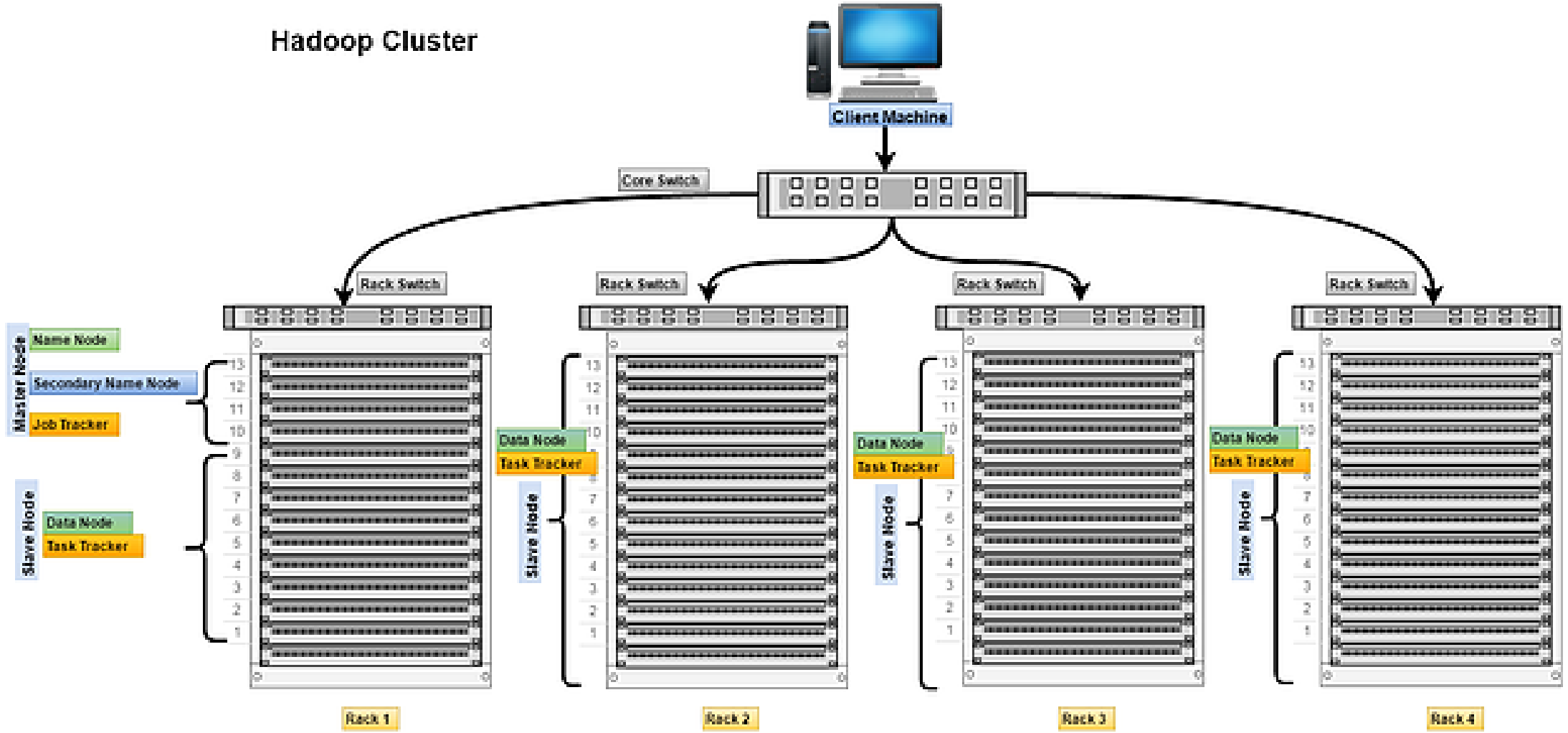
Secondary name node

- Przechowuje logi replikowane w określonym czasie z *name node*
- Zadaniem *secondary name node* jest redukcja czasu zarządzania metadanymi klastra oraz czasu restartu klastra
- *secondary name node stanowi* zapisuje stany danych(*checkpoint*) w systemie HDFS; służy to wsparciu wydajności pracy *name node*
- *secondary name node* nie służy zapewnieniu wysokiej dostępności klastra (HA)

Data node

- Składuje dane na systemie HDFS
- Przekazuje informacje od *name node*, dotyczące swojego statusu oraz posiadanych plików
- Może pracować w trybie replikacji danych (także RAID)
- Wykonuje zadania obliczeniowe zlecane poprzez MapReduce lub Yarn

Architektura fizyczna

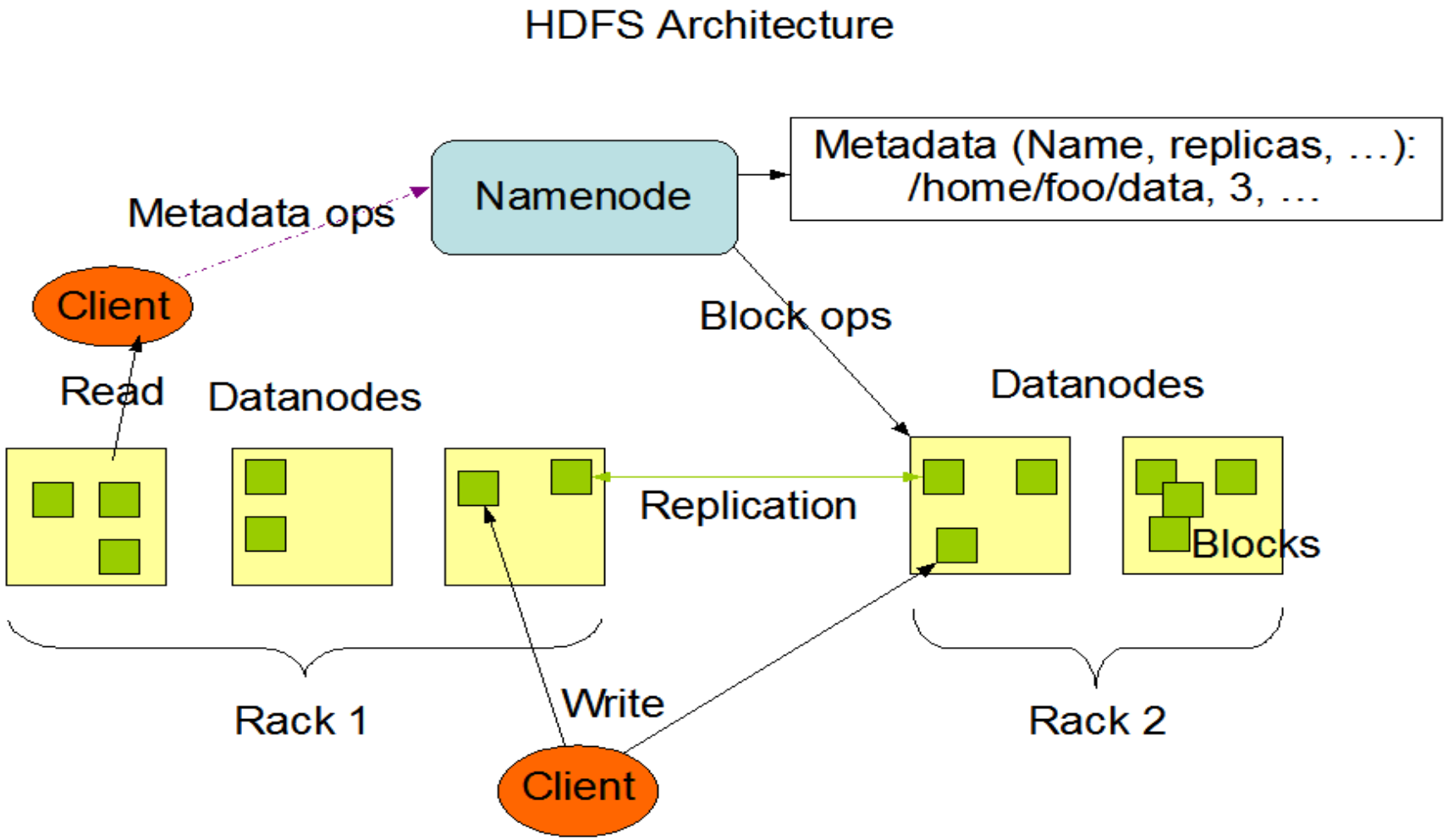


Źródło: pacificbigdata.com

Architektura fizyczna

Źródło: pacificbigdata.com

Apache Hadoop: zasady zapisu/odczytu



Źródło: hadoop.apache.org

Kiedy nie stosować HDFS

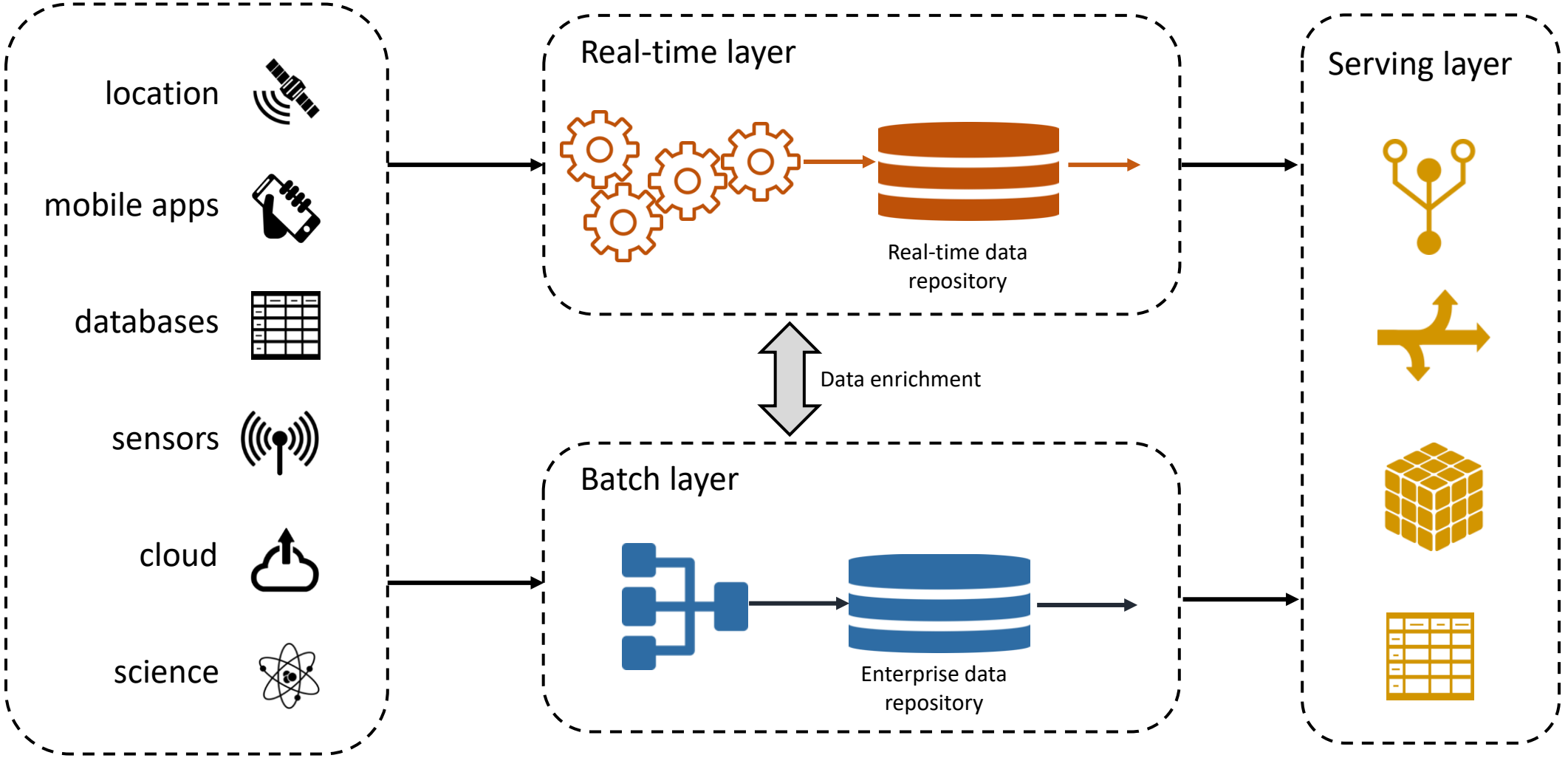
- *Low latency* and *real time*
- Dane posiadają strukturę
- Wolumen nie jest bardzo duży
- Dużo zapisów (więcej niż odczytów)
- Jeśli algorytm nie daje się dekomponować na równoległe kroki

Nowoczesne architektury

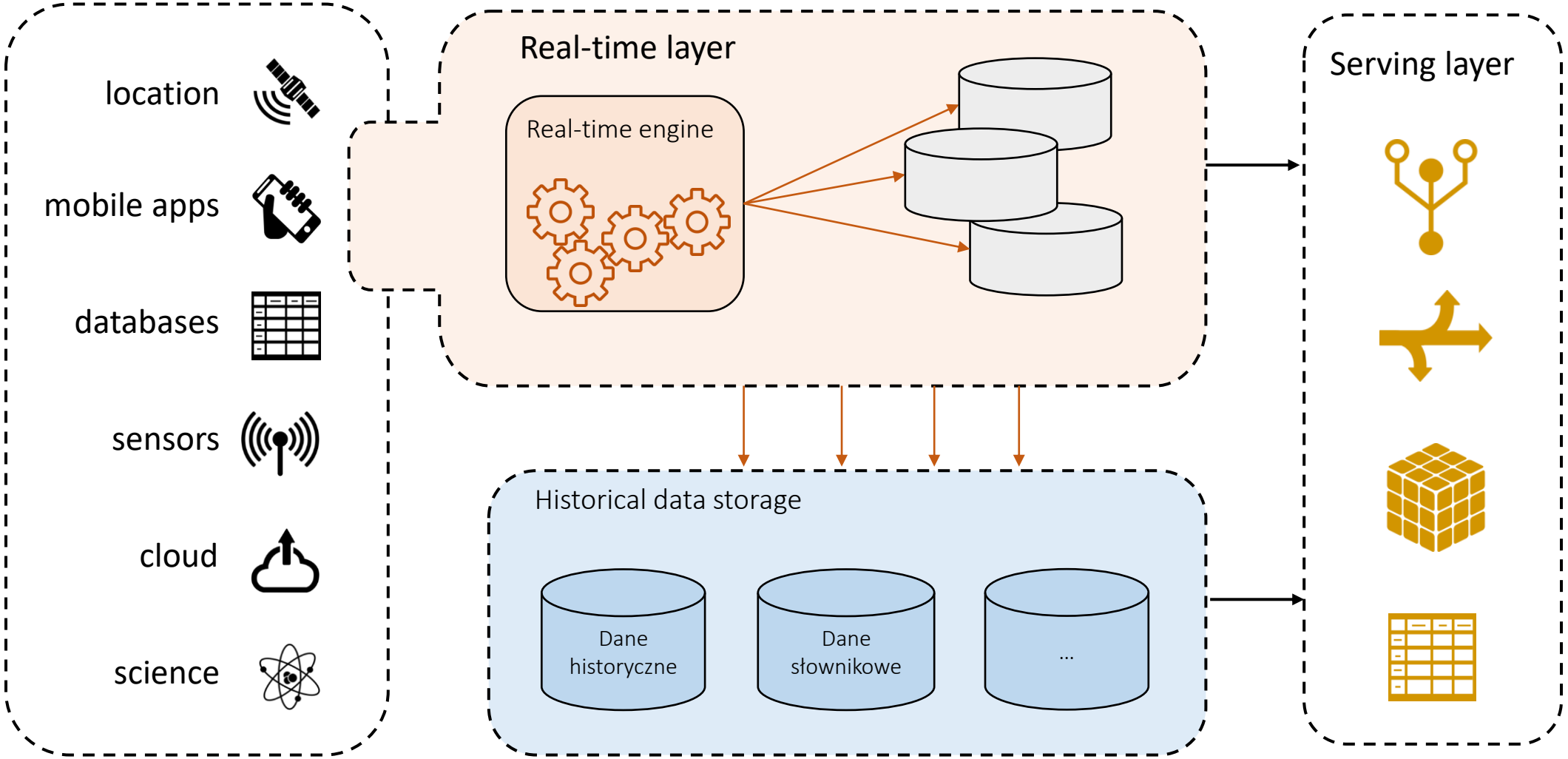
Data Lake

- Repozytorium służące składowaniu i przetwarzaniu danych o bardzo dużej skali i zróżnicowaniu
- Możliwość podłączania zróżnicowanych źródeł danych, zarówno posiadających strukturę jak i pozbawionych struktury; danych wsadowych oraz strumieni
- Dane nie są składowane w sposób uporządkowany jak w przypadku hurtowni danych czy data martów. Jest to często federacja technologii, baz danych i strumieni danych
- Architektura powstała jako odpowiedź na wady „klasycznych” hurtowni danych:
  - HD odpowiadają tylko na pytania, które były znane wcześniej
  - Hurtownie danych i data marty posiadają dane o określonej szczegółowości. Nie można jej zwiększyć
  - HD opierają się na zdefiniowanych źródłach danych

Architektura *lambda*



Architektura *kappa*



Databricks

https://community.cloud.databricks.com

Dziękuję za uwagę