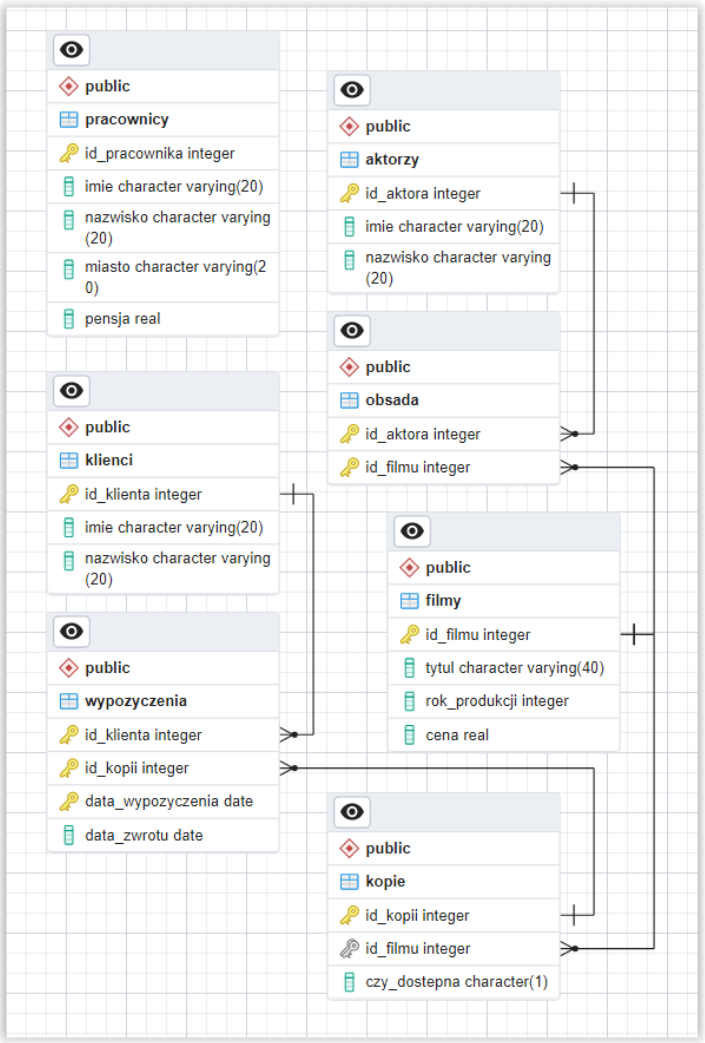
Przetwarzanie strumieni danych w systemach Big Data

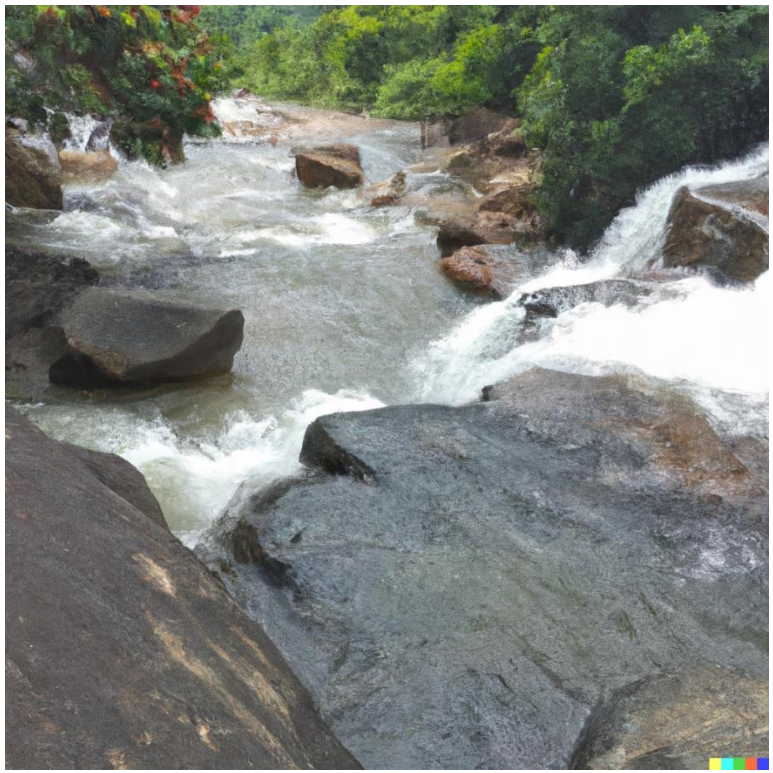
wprowadzenie

Krzysztof Jankiewicz

Jaki jest świat danych?



Wsadowy czy strumieniowy?



Dlaczego?

Już w 2015 roku Databricks przeprowadziła badanie wśród swoich użytkowników dotycząca wykorzystywania przez nich mechanizmów Spark Streaming
Okazało się, że 56% z nich korzysta z nich, a 48% uważa, że jest kluczowy element w ich biznesie.

Dlaczego?

czyli jak slajd nie powinien wyglądać

- ***Real-time analytics***: Przetwarzanie strumieni danych umożliwia analizę danych w czasie rzeczywistym, co umożliwia przedsiębiorstwom natychmiastowe reagowanie na zmiany w danych i podejmowanie lepszych decyzji biznesowych. Dzięki temu firmy mogą szybciej dostosowywać swoje strategie i podejmować odpowiednie działania, co przekłada się na poprawę wyników finansowych.
- ***Efektywność operacyjna***: Przetwarzanie strumieni danych pozwala na automatyzację procesów biznesowych i eliminację opóźnień w przetwarzaniu danych, co przekłada się na znaczne usprawnienie działalności przedsiębiorstw. Dzięki temu zespoły mają więcej czasu na podejmowanie decyzji biznesowych i rozwijanie strategii.
- ***Zwiększenie konkurencyjności***: Przetwarzanie strumieni danych umożliwia przedsiębiorstwom szybkie i efektywne wykorzystanie danych do podejmowania decyzji biznesowych. Dzięki temu firmy mogą lepiej dostosować się do zmieniającego się otoczenia rynkowego i zwiększyć swoją konkurencyjność.
- ***Ulepszona jakość usług***: Przetwarzanie strumieni danych umożliwia przedsiębiorstwom monitorowanie jakości swoich usług w czasie rzeczywistym i natychmiastowe reagowanie na sytuacje awaryjne. Dzięki temu przedsiębiorstwa mogą zwiększyć zadowolenie klientów poprzez zapewnienie im lepszej jakości usług.
- ***Redukcja kosztów***: Przetwarzanie strumieniowe strumieni danych przedsiębiorstwom na automatyzację procesów biznesowych i eliminację opóźnień w przetwarzaniu danych, co przekłada się na redukcję kosztów operacyjnych. Dodatkowo, dzięki szybszemu dostępowi do informacji, firmy mogą podejmować decyzje biznesowe na podstawie bardziej aktualnych danych, co pozwala uniknąć kosztownych błędów.
- ***Możliwość szybkiego reagowania na problemy***: Przetwarzanie strumieni danych pozwala na szybkie wykrycie problemów w procesach biznesowych i natychmiastowe podjęcie działań mających na celu ich rozwiązanie. Dzięki temu przedsiębiorstwa mogą uniknąć kosztownych przerw w działalności i utrzymanie ciągłości biznesowej.

Przetwarzanie Strumieni Danych

- What,
- Where,
- When, and
- How

  of Large-Scale Data Processing

Streaming Systems

Tyler Akidau, Slava Chernyak, Reuven Lax

O'Reilly Media, II wydanie, 2019

Wstęp

- Czym jest strumień?
- Czym jest silnik/system przetwarzania strumieni danych?
- Aplikacje przetwarzające strumienie danych

Czym jest strumień?

- Dwa wymiary pojęcia: liczność i natura
- Liczność
  - Dane ograniczone (*bounded data*) – zbiór danych, który ma skończony rozmiar
  - **Dane nieograniczone** (*unbounded data*) – zbiór danych, który ma nieskończony (teoretycznie) rozmiar
- Natura (charakter)
  - Tabela – zbiór danych z określonego momentu w czasie
  - **Strumień** – dane, które prezentują, element po elemencie, ewolucję (zmianę) danych w czasie.

Przykłady strumieniowych źródeł danych

- Transakcje giełdowe w czasie rzeczywistym
- Zarządzanie zapasami w handlu
- Aplikacje do udostępniania przejazdów
- Gry dla wielu graczy
- Internet of Things
- Systemy śledzenia lokalizacji
- Transakcje bankowe

Czym jest silnik/system przetwarzania strumieni danych?

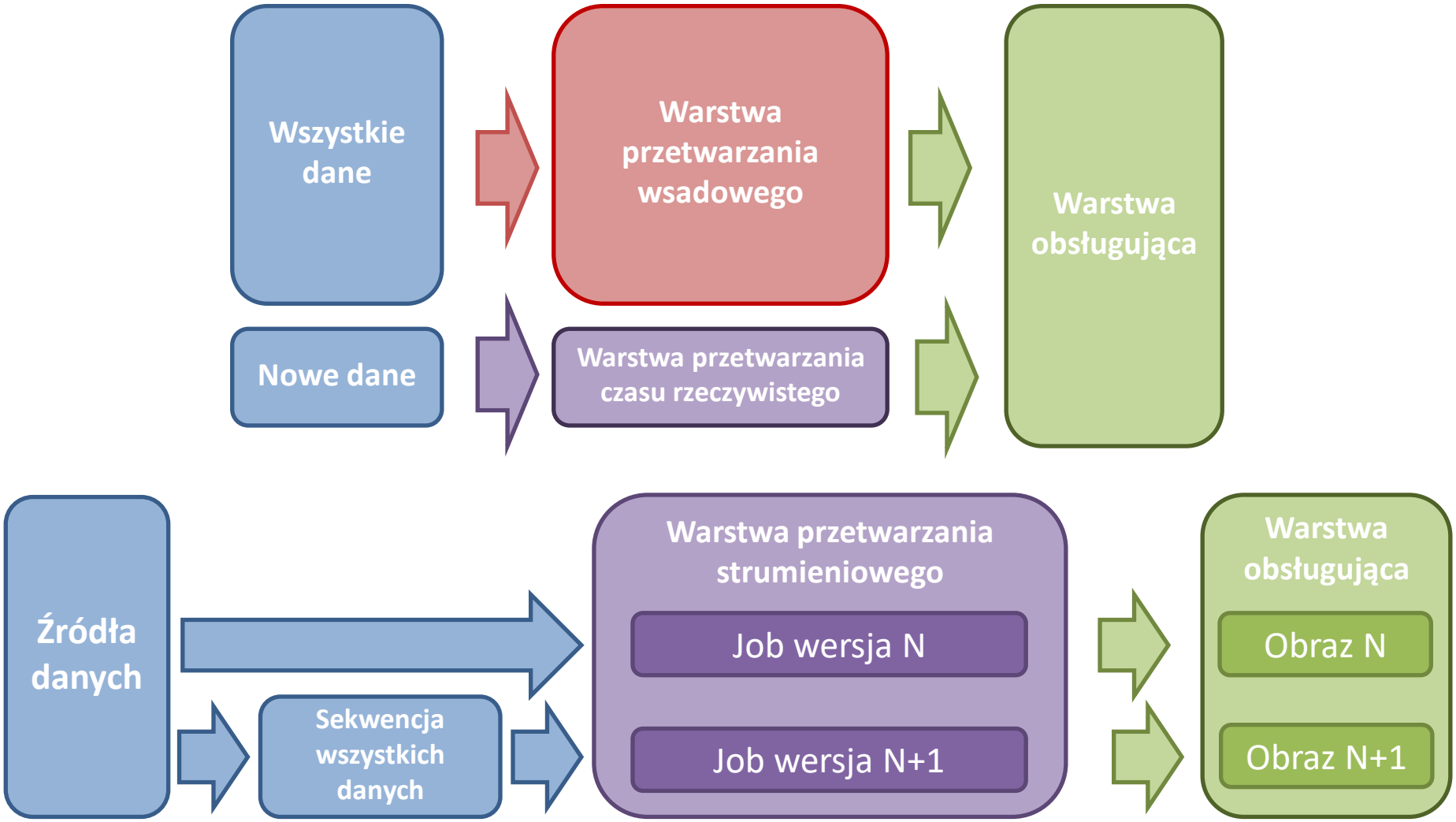
Typ silnika przetwarzania danych zaprojektowany z myślą o nieskończonych zbiorach danych

Przetwarzanie *Batch, Micro-Batch, Stream*

- Skończone zbiory danych – przetwarzanie wsadowe.

- Nieskończone zbiory danych
  - Przetwarzanie mikro-wsadowe
    - stałe okna – problem kompletności danych
    - okna sesyjne – rozbicie danych pomiędzy mikro-wsadami
  - Przetwarzanie strumieni danych (ciągłe/pełnowymiarowe)

Architektury systemów Big Data – przypomnienie



Ewolucja systemów przetwarzania strumieni danych

Początki w ramach projektów badawczych, ale także komercyjnych 1990

Generacje rozproszonych systemów przetwarzania strumieni danych:

- Pierwsza generacja (2011)
  - małe opóźnienia
  - niskopoziomowe API
  - **brak obsługi etykiet zdarzeń** – brak powtarzalności, spójności i dokładności wyników
  - gwarancje "at-least once"
  - wykorzystanie w architekturze Lambda
- Druga generacja (2013)
  - API wysokiego poziomu
  - lepsza obsługa awarii
  - zwiększona przepustowość
  - zwiększone opóźnienia
  - nadal oparcie się na czasie i kolejności przybywania zdarzeń
- Trzecia generacja (2015)
  - **wykorzystanie etykiet zdarzeń**
  - gwarancje "exactly-once"
  - możliwość konfigurowania przepustowości/opóźnienia
  - możliwość obsługi danych bieżących oraz historycznych
  - wyniki powtarzalne, spójne i dokładne
  - możliwe wykorzystanie architektury Kappa

Hueske, Fabian; Kalavri, Vasiliki.
Stream Processing with Apache Flink; O'Reilly Media

Czym się charakteryzują systemy przetwarzania strumieni danych (*Stream Data Processing*)?

- Odbierają dane w sposób ciągły
- Działają 24/7 – duża waga mechanizmów obsługi awarii
- Przetwarzają dane na bieżąco
- Wyniki dostępne są z tzw. niską latencją
- Miejsca docelowe o dodatkowej funkcjonalności
  - Duża częstotliwość operacji
  - Możliwa potrzeba aktualizacji danych (a nie tylko dopisywania nowych)
- Możliwe ograniczenia dotyczące dokładności wyniku, stosowania przybliżeń, heurystyk
- Stosunkowo mniejsza przepustowość (rozłożona w czasie)

Aplikacje przetwarzające strumienie danych
Przypadki zastosowań

- Analiza lokalizacji
- Wykrywanie oszustw
- Transakcje giełdowe w czasie rzeczywistym
- Marketing, sprzedaż i analityka biznesowa
- Aktywność klienta / użytkownika (aplikacji, portalu, urządzeń)
- Monitorowanie i raportowanie wewnętrznych systemów informatycznych
- Monitorowanie dzienników: rozwiązywanie problemów z systemami, serwerami, urządzeniami i nie tylko
- Uczenie maszynowe i sztuczna inteligencja: łączenie przeszłych i obecnych danych
- SIEM (Security Information and Event Management): analizowanie dzienników i danych o zdarzeniach w czasie rzeczywistym w celu monitorowania, pomiarów i wykrywania zagrożeń
- Zapasy w handlu detalicznym / magazynie: zarządzanie zapasami we wszystkich kanałach i lokalizacjach oraz zapewnianie bezproblemowej obsługi na wszystkich urządzeniach
- Zarządzanie pojazdami współdzielonymi: łączenie danych dotyczących lokalizacji, użytkownika i cen na potrzeby analiz predykcyjnych; dopasowywanie kierowców do najlepszych kierowców pod względem bliskości, miejsca docelowego, cen i czasu oczekiwania

Za: `https://www.confluent.io/learn/data-streaming/`

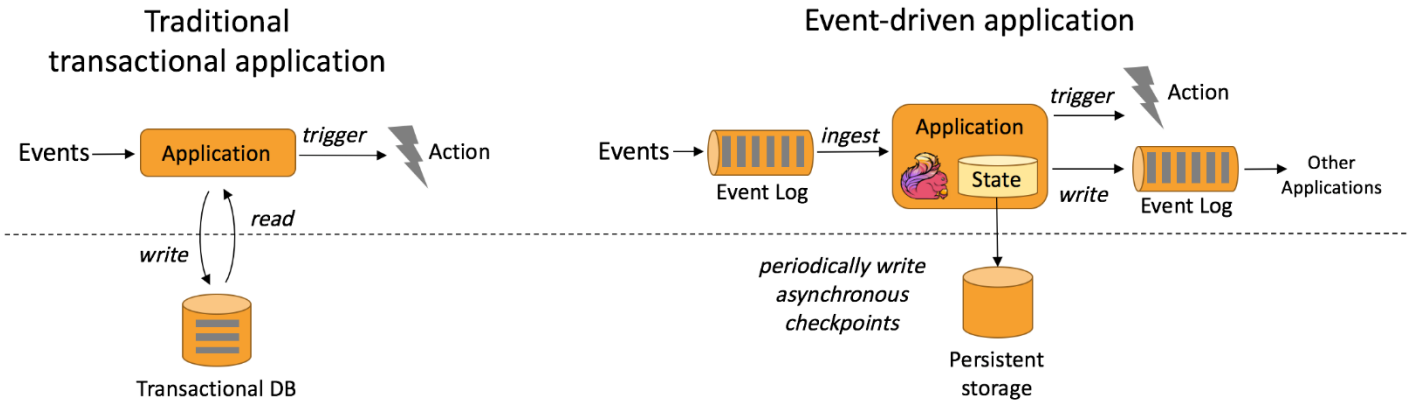
Główne przykłady zastosowań

- Aplikacje oparte na zdarzeniach (rekomendacje czasu rzeczywistego, wykrywanie wzorców, CEP, wykrywanie anomalii)
- Przepływy danych
- Analityka strumieni danych (monitorowanie jakości, zachowania użytkowników)

Aplikacje oparte na zdarzeniach

Event-driven Applications

- Cechy
  - odczyt z jednego lub wielu źródeł strumienia zdarzeń
  - natychmiastowa reakcja na pojawiające się zdarzenia
- Rozwiązania klasyczne vs strumieniowe



- Zalety
  - znacznie większa przepustowość
  - mniejsze opóźnienia

`https://flink.apache.org/usecases.html`

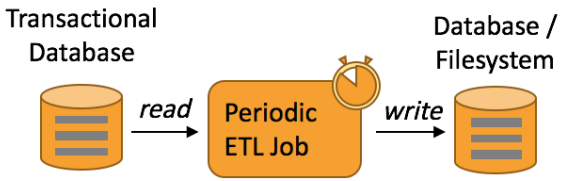
- Przykłady:
  - detekcja anomalii
  - detekcja oszustw
  - alarmy, których definicje oparte są na regułach
  - monitorowanie przetwarzania procesów biznesowych
  - aplikacje internetowe (analiza ruchu, aktywności użytkowników)

Przepływy danych

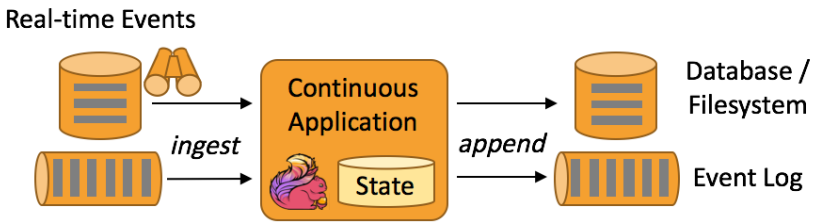
Data Pipeline Applications

- Cechy
  - odpowiednik operacji ETL
  - transformacja i przenoszenie danych realizowane w sposób ciągły
- Rozwiązania klasyczne vs strumieniowe

Periodic ETL



Data Pipeline



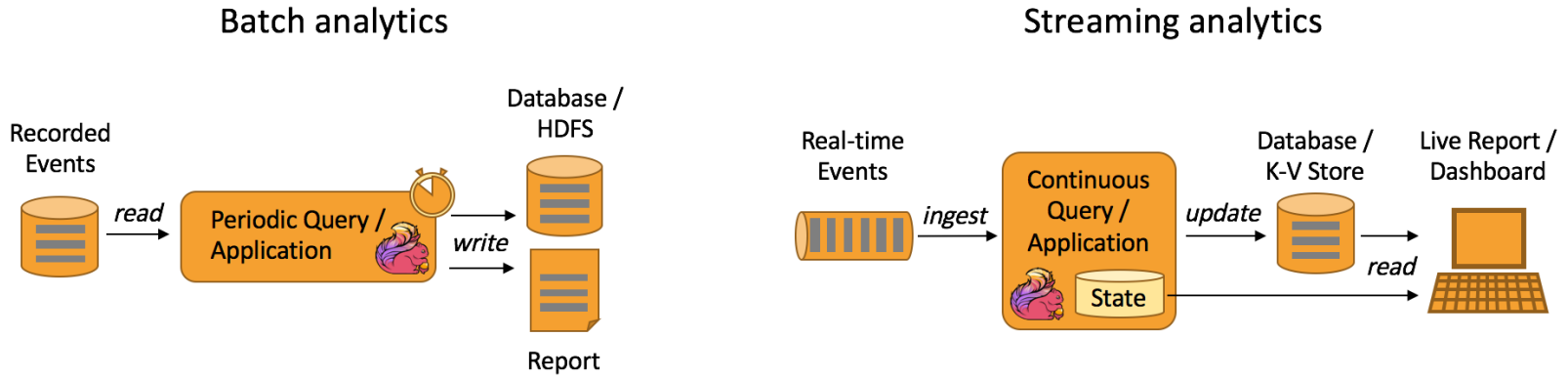
- Zalety
  - mniejsze opóźnienia
  - znacznie większa liczba przypadków użycia
- Główne obszary zastosowań
  - tworzenie indeksów dla silników wyszukiwani
  - ETL czasu rzeczywistego

`https://flink.apache.org/usecases.html`

Analityka strumieni danych

Data Analytics Applications

- Cechy
  - analiza danych pochodzących ze źródeł
  - wyniki analiz są zapisywane lub bezpośrednio prezentowane
- Rozwiązania klasyczne vs strumieniowe



https://flink.apache.org/usecases.html

- Zalety:
  - zmniejszone opóźnienia
  - prostsza architektura
- Przykłady zastosowań
  - Monitorowanie jakości usług (np. sieci telekomunikacyjnych)
  - Analiza danych grafowych dużej skali
  - Analiza danych szybko zmieniających się (np.: zachowania użytkowników, współpracowników, maszyn)

Pytania?