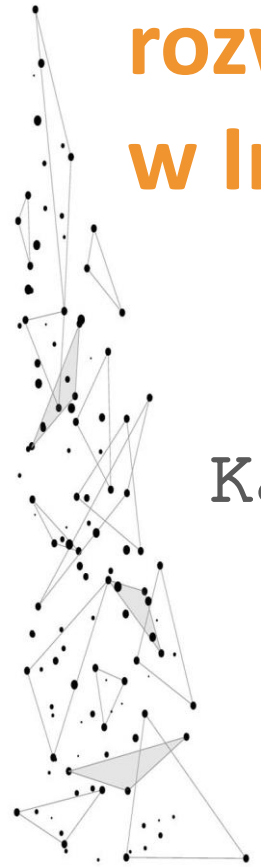
Developers KATOWICE

FP Data Solutions

Apache Spark on Azure

rozwiązanie do przetwarzania danych w Internet of Things



Kamil Dworak

FP Data Solutions

O mnie

Kamil Dworak

FP Data Solutions

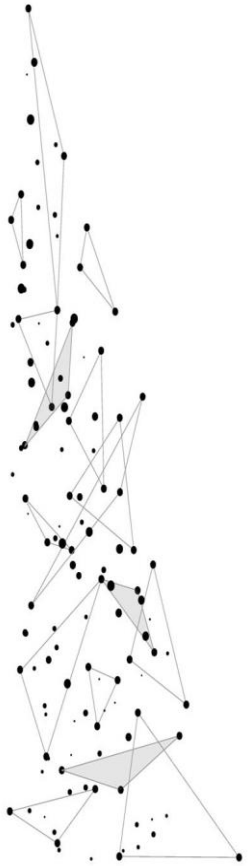
UNIWERSYTET ŚLĄSKI
W KATOWICACH

https://github.com/devkam/Events/

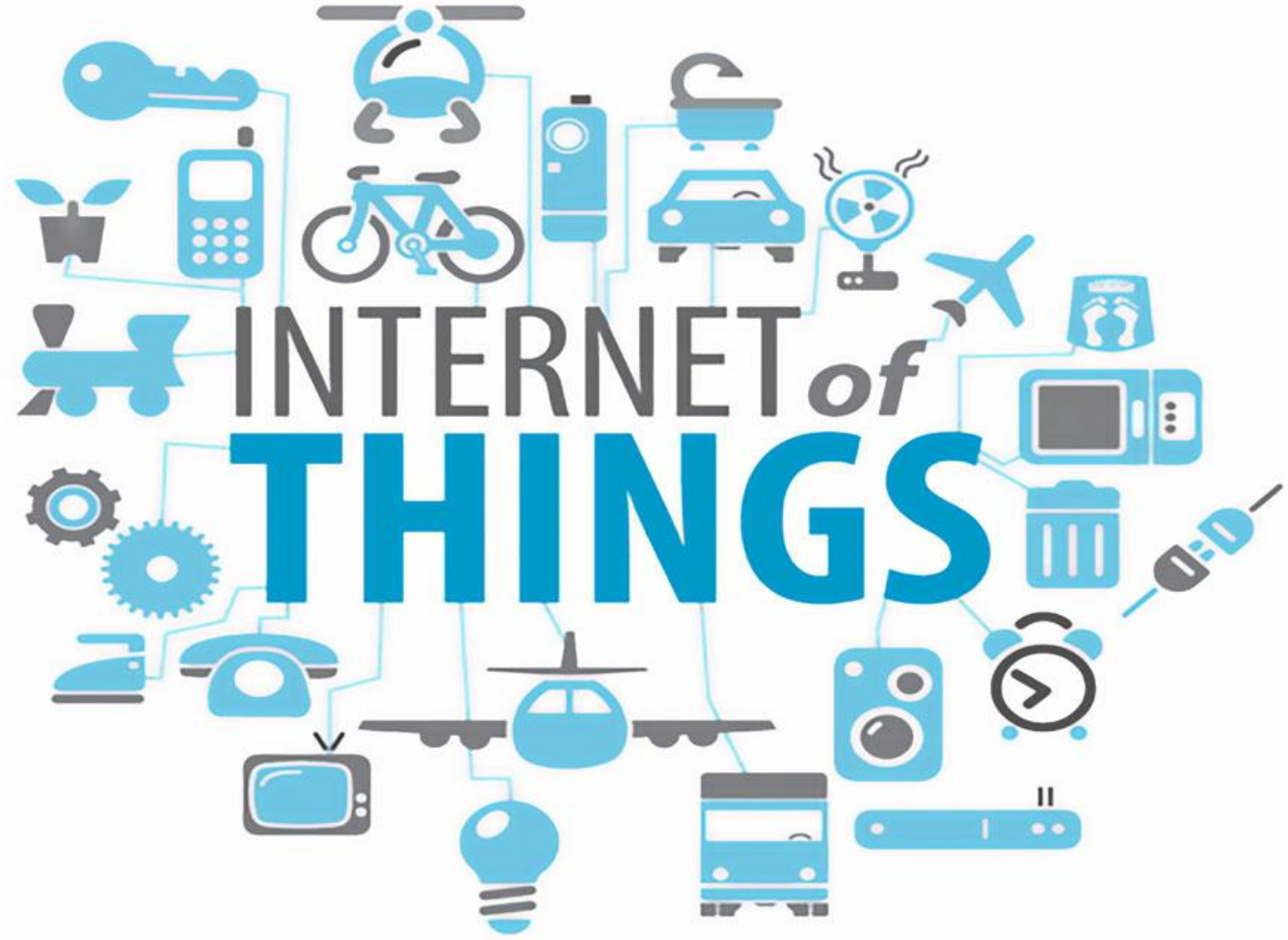
FP Data Solutions

Agenda

- Internet of Things
  - Stream Data Processing
  - Apache Kafka
- Dlaczego chmura?
- Azure jako platforma dla danych near to realtime
  - Azure Event Hubs oraz Azure IoT Hubs
  - Azure Databricks
- Spark Structure Streaming
  - Demo
- Q&A



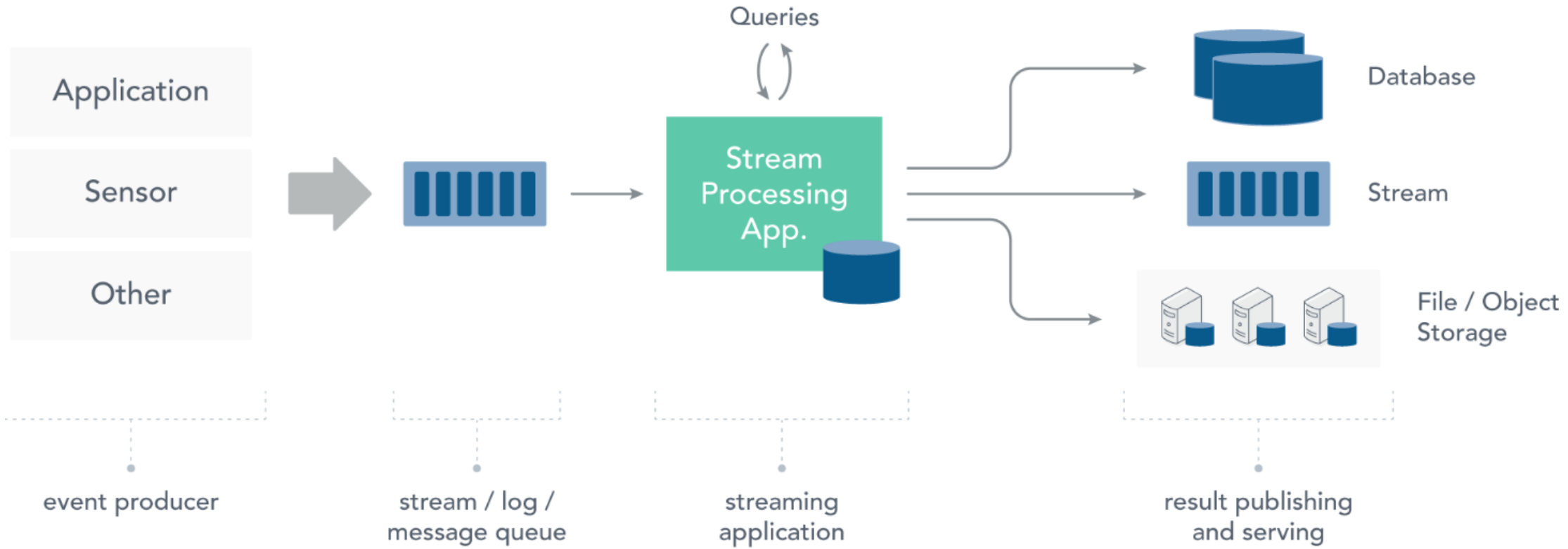
FP Data Solutions



Źródło: www.com-strat.com

FP Data Solutions

Stream Data Processing



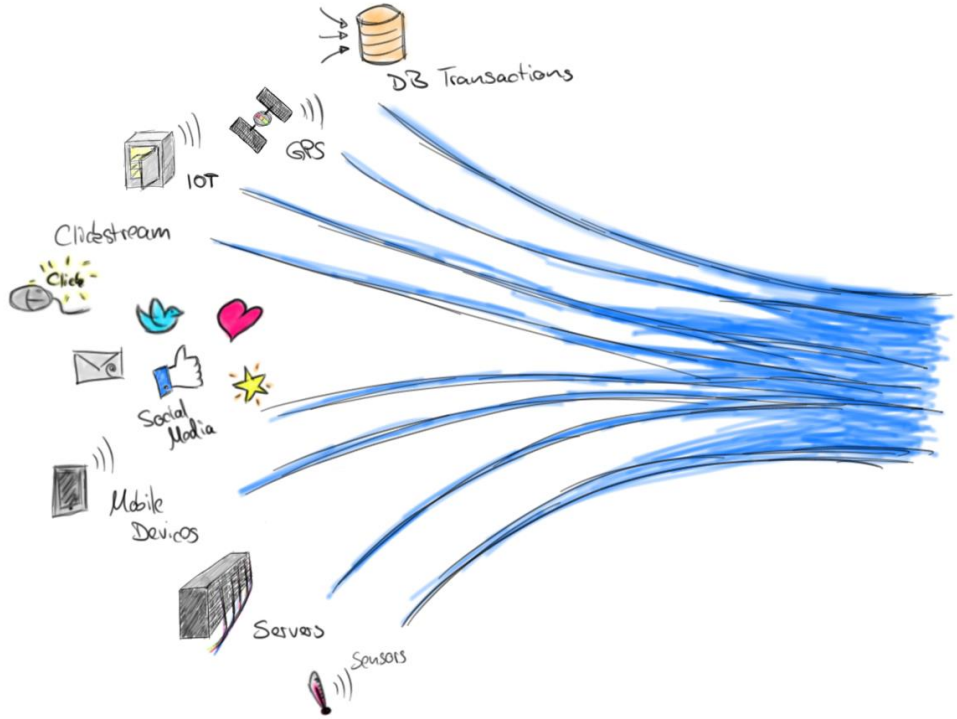
Stream Processing for Batch/Realtime Data Processing, and Event-driven Applications

Źródło: www.ververica.com

FP Data Solutions

Stream Data Processing

- Teoretycznie nieskończony strumień z danymi
- Dostęp do danych jest niemal natychmiastowy
- Przetwarzanie musi odbywać się tak szybko, jak to tylko możliwe
- Stan i rozmiar danych jest nieznany

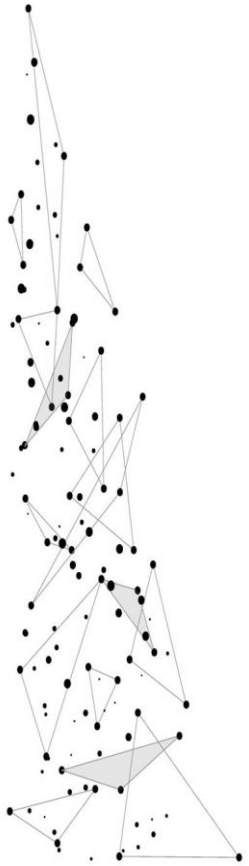


Źródło: flink.apache.org

FP Data Solutions

Apache Kafka

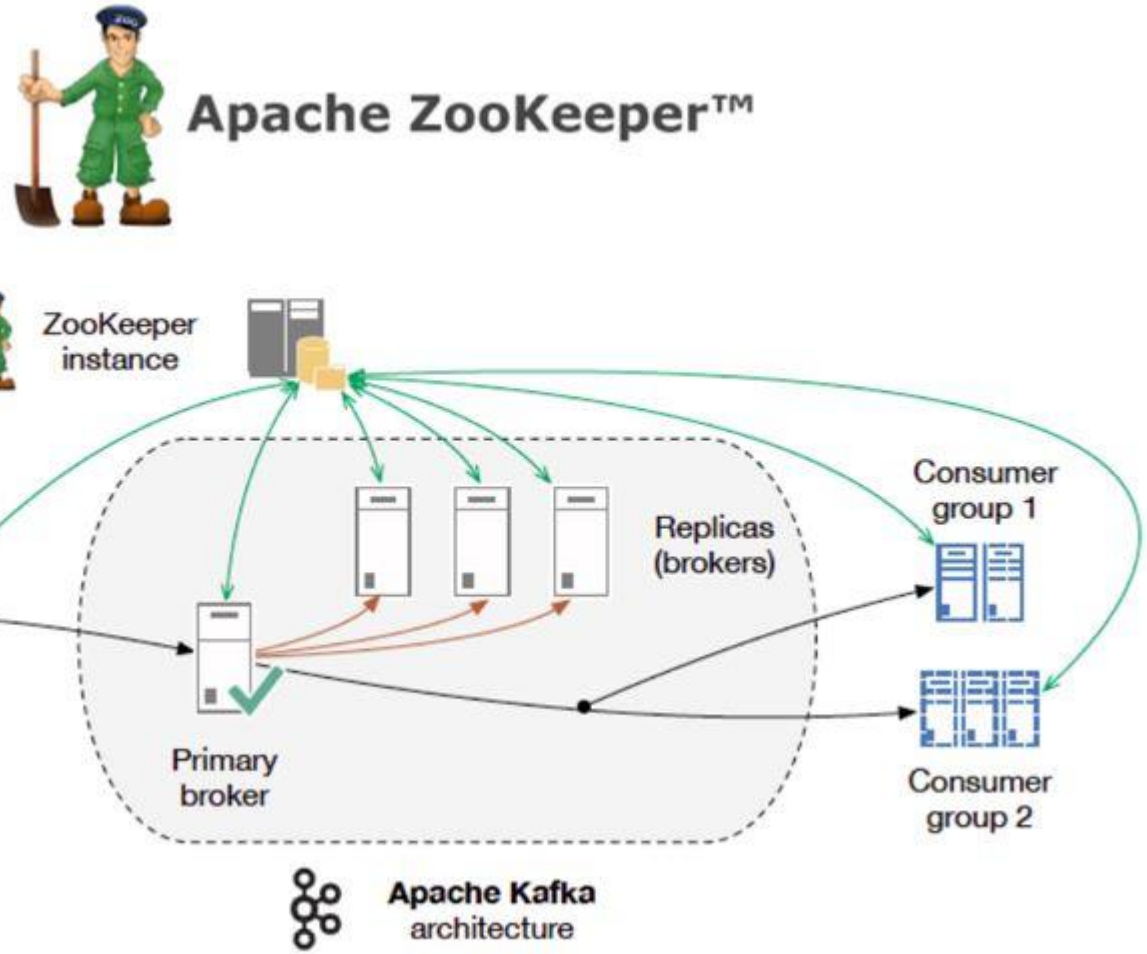
- Rozproszony i skalowalny message broker
- Zastosowanie w Big Data
- Gwarantuje dostarczenie wiadomości
- Model publisher-subscriber
- Pozwala na pracę w trybie batch’owym



kafka

FP Data Solutions

Apache Kafka



Źródło:
www.prathapkudupublog.com

FP Data Solutions

Use Case

Źródło: esdnews.com.au

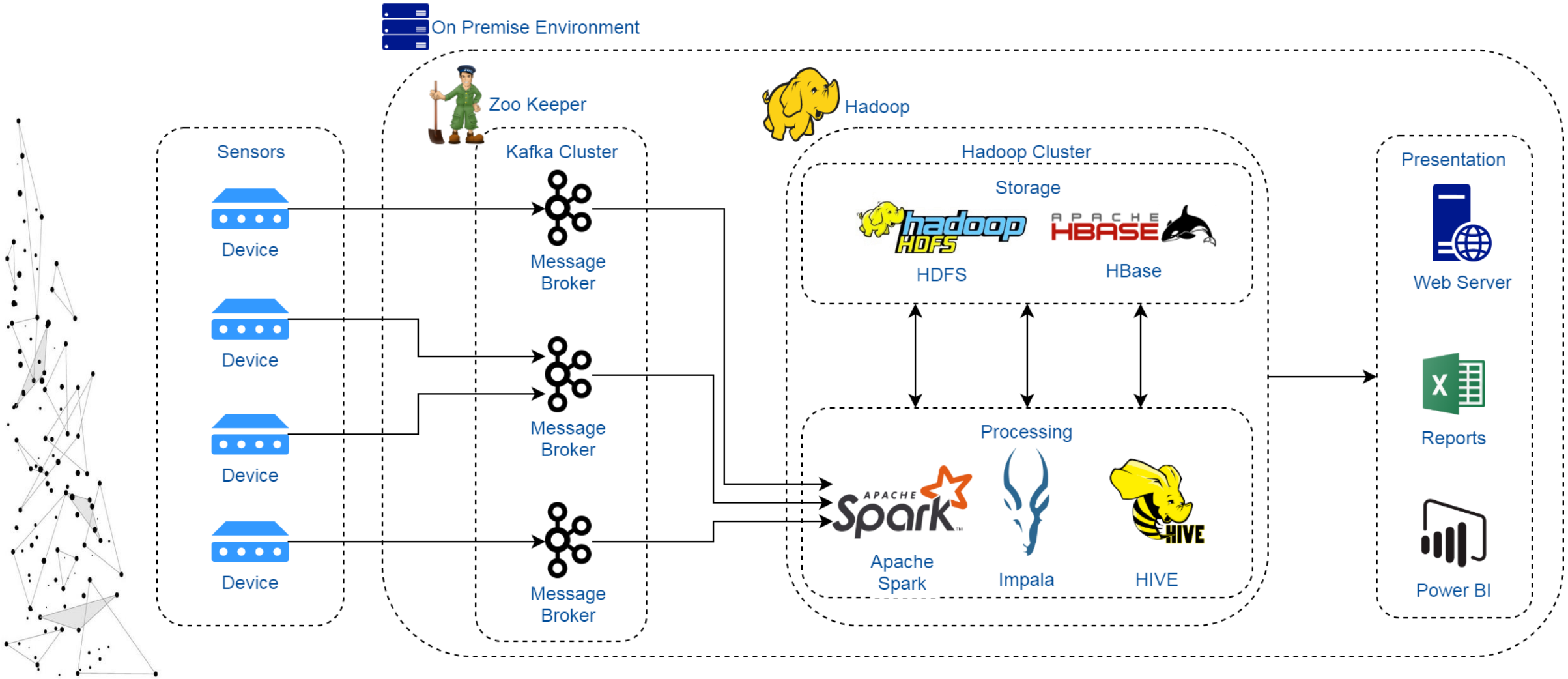
FP Data Solutions

Co nam będzie potrzebne?

- Transfer danych z urządzeń
- Obszerny magazyn na dane (TB – PB)
- Przetwarzanie dużych zbiorów danych
  - Wybrane algorytmy do analizy danych pomiarowych
- Skalowalność

FP Data Solutions

Jakby to mogło wyglądać na on premise?



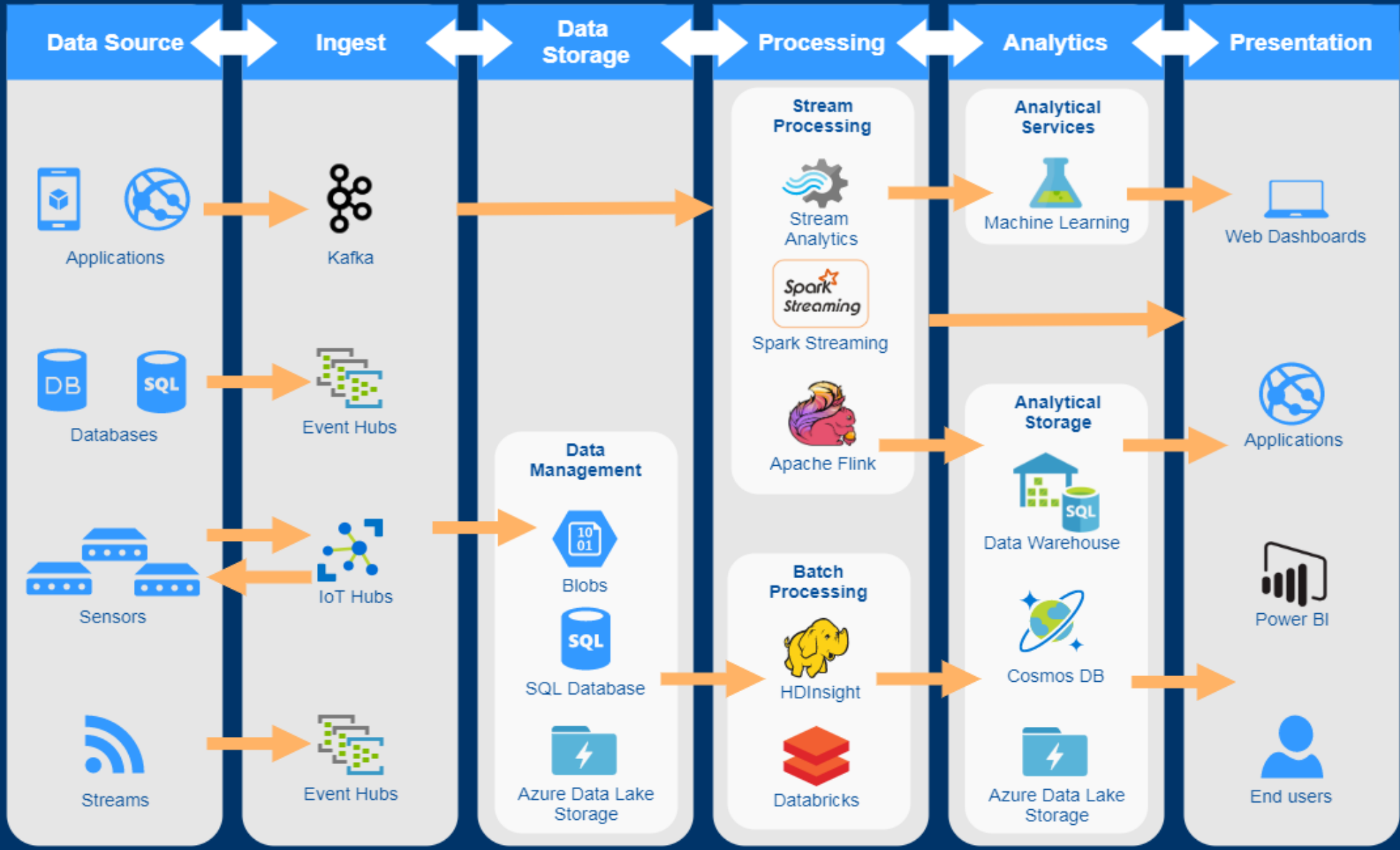
FP Data Solutions

Co pociąga za sobą on premise?

- Duży wolumen danych
- Instalacja, konfiguracja, utrzymanie
- Nieznane koszty
- Skalowalność
- Odpowiedzialność za bezpieczeństwo systemu

computing © Jakub Jirsák - Fotolia.com

Stream Data Processing



FP Data Solutions

Co przeważa za cloudem ?

- Nieograniczony storage
- Nieograniczona moc obliczeniowa
- Produktywność
- Dużo niższy próg wejścia
- Wysoka skalowalność i wydajność
- Poczucie bezpieczeństwa (SLA)
- Bezpieczeństwo danych
  - Replikacja danych

amazon web services

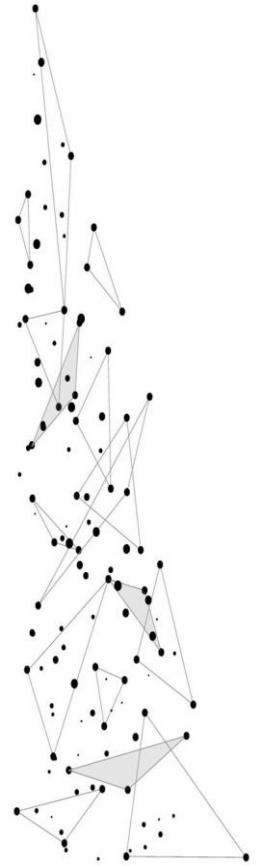
Azure

Google Cloud Platform

FP Data Solutions

Azure Event Hub

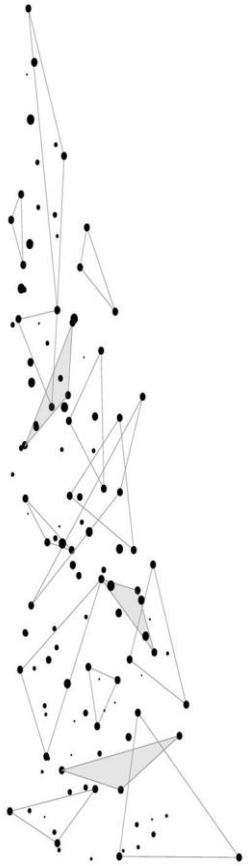
- Pozwala na pozyskanie milionów wiadomości na sek.
- Jest w pełni skalowalny
- Posiada wsparcie dla protokołów Kafki
- Wspiera AQMP, HTTPS
- Event Capturing
- Dane przechowywane są za pośrednictwem partycji



FP Data Solutions

Azure IoT Hub

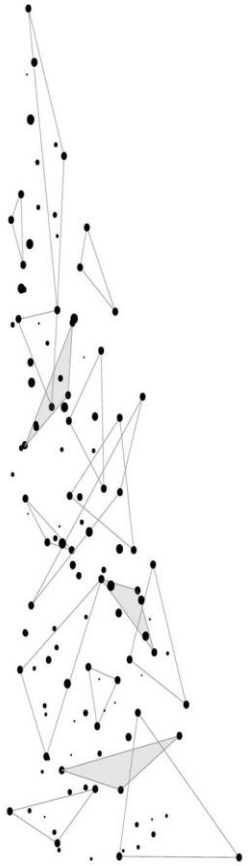
- Dwukierunkowa komunikacja
- Wsparcie dla protokołu MQTT
- Możliwość wysyłania plików
- Konfiguracja indywidualnych tożsamości



FP Data Solutions

Azure Databricks

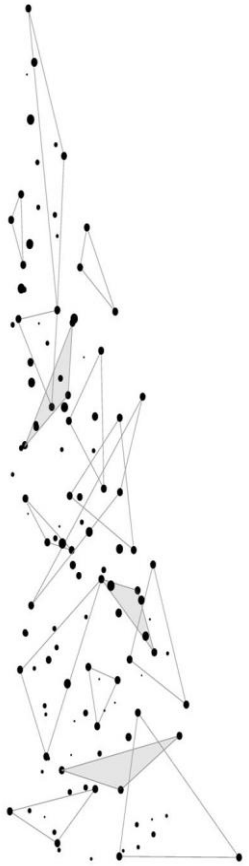
- Zoptymalizowane środowisko Apache Spark (CaaS)
- Automatyczne skalowanie
- Prosta integracja z innymi usługami PaaS'owych
- Pozwala na analizę w czasie rzeczywistym
- Wbudowane algorytmy uczenia maszynowego
- Wsparcie wielu języków programowania



FP Data Solutions

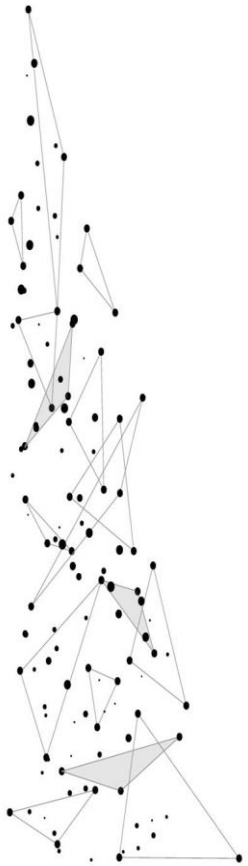
Spark Structured Streaming

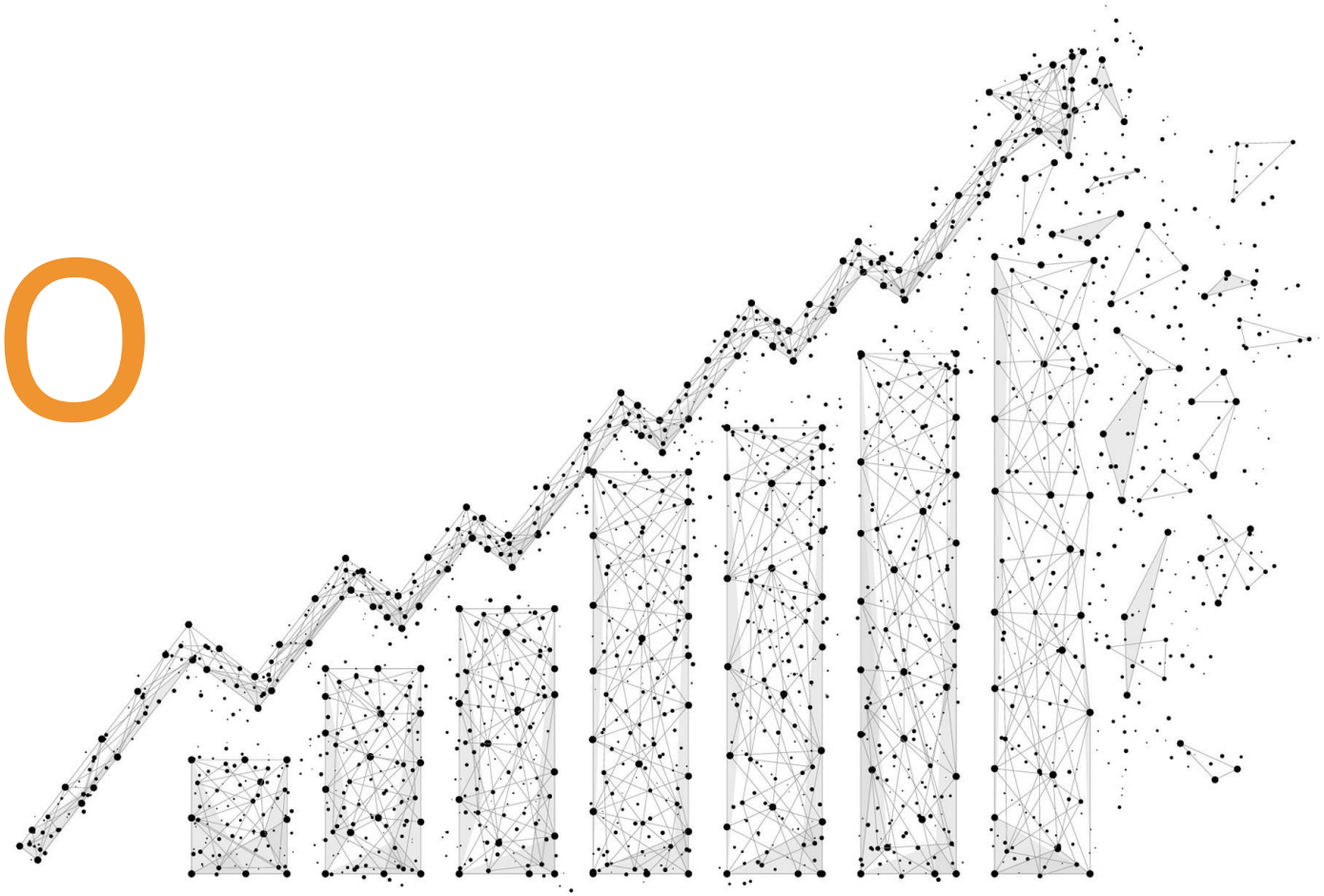
- Komponent Sparka odpowiedzialny na przetwarzanie strumieniowe
- Zgodny z Dataframe API
- Działa o 10-20x szybciej niż v. 1.0
- Łatwo skalowalny silnik przetwarzania danych o charakterze near to real time



FP Data Solutions

DEMO





www.future-processing.com

FP Data Solutions

Wnioski

- Cloud stanowi bardzo dobre środowisko dla projektów Big Data o ruchomej skali problemu
- Chmura pozwala na o wiele szybszy start
- Kafka doskonale odnajduje się w projektach o charakterze Big Data
- Storage jest tani, usługi związane z przetwarzaniem danych są drogie
- Nie zawsze musimy brać za wszystko pełną odpowiedzialność
- **Vendor lock-in !**

FP Data Solutions

Resources

- https://kafka.apache.org/documentation/
- https://docs.microsoft.com/pl-pl/azure/event-hubs/
- https://azure.microsoft.com/pl-pl/services/databricks/
- https://docs.databricks.com/
- https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html

FP Data Solutions

CLOUDYNA 2019

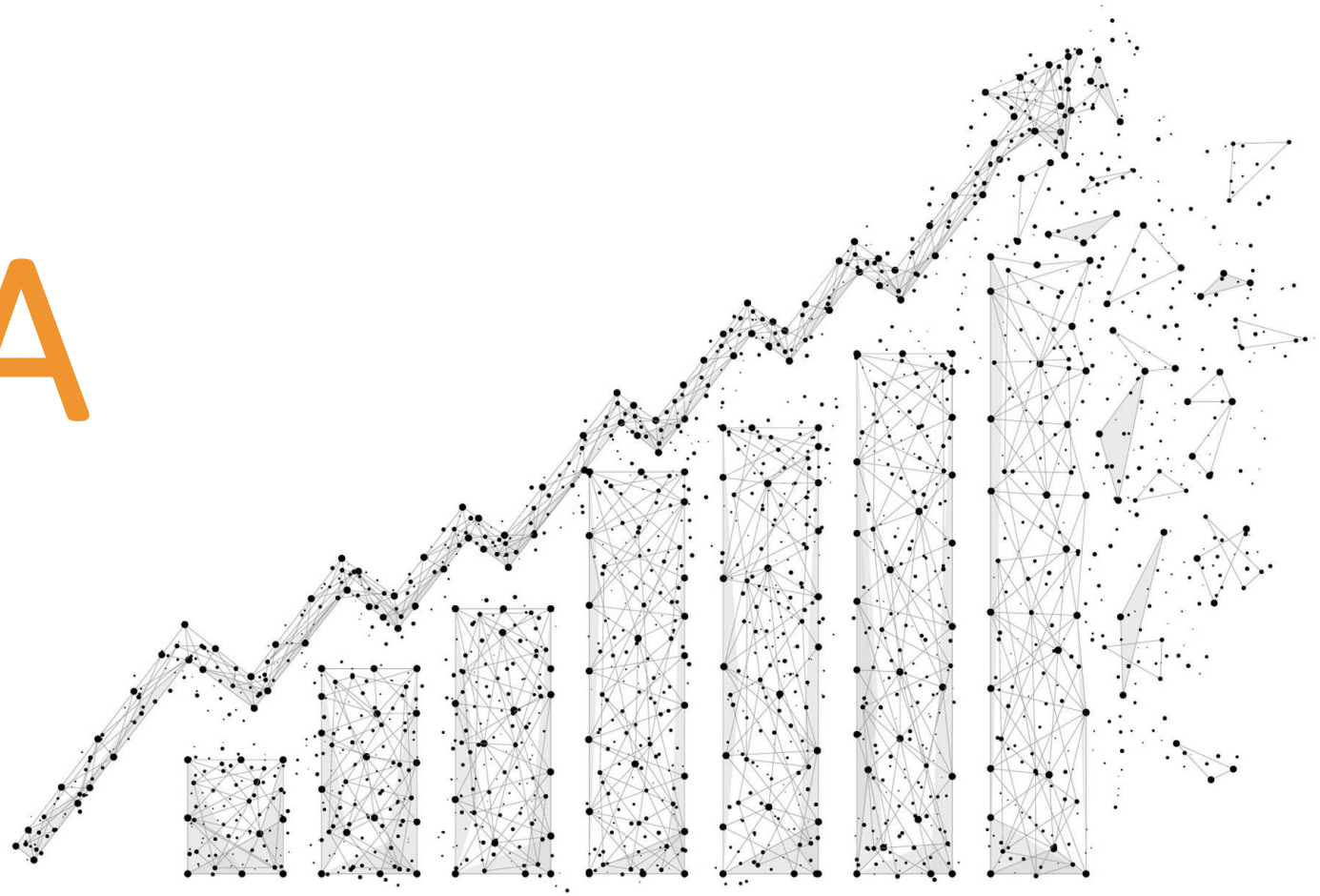
PUBLIC CLOUD FEST IN KATOWICE

November 13th 2019

Katowice, Poland

FP Data Solutions

Q & A



www.future-processing.com

FP Data Solutions

THANK YOU!

www.future-processing.com