GLIWICE SOFTWARE BARCAMP

FP Data Solutions

Jak zbudować rozwiązanie przetwarzające terabajty danych pochodzące z inteligentnych liczników w skończonym czasie

Kamil Dworak

www.future-processing.com

FP Data Solutions

O mnie

Kamil Dworak

FP Data Solutions

UNIWERSYTET ŚLĄSKI
W KATOWICACH

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

http://bujacwoblokach.pl/

FP Data Solutions

Agenda

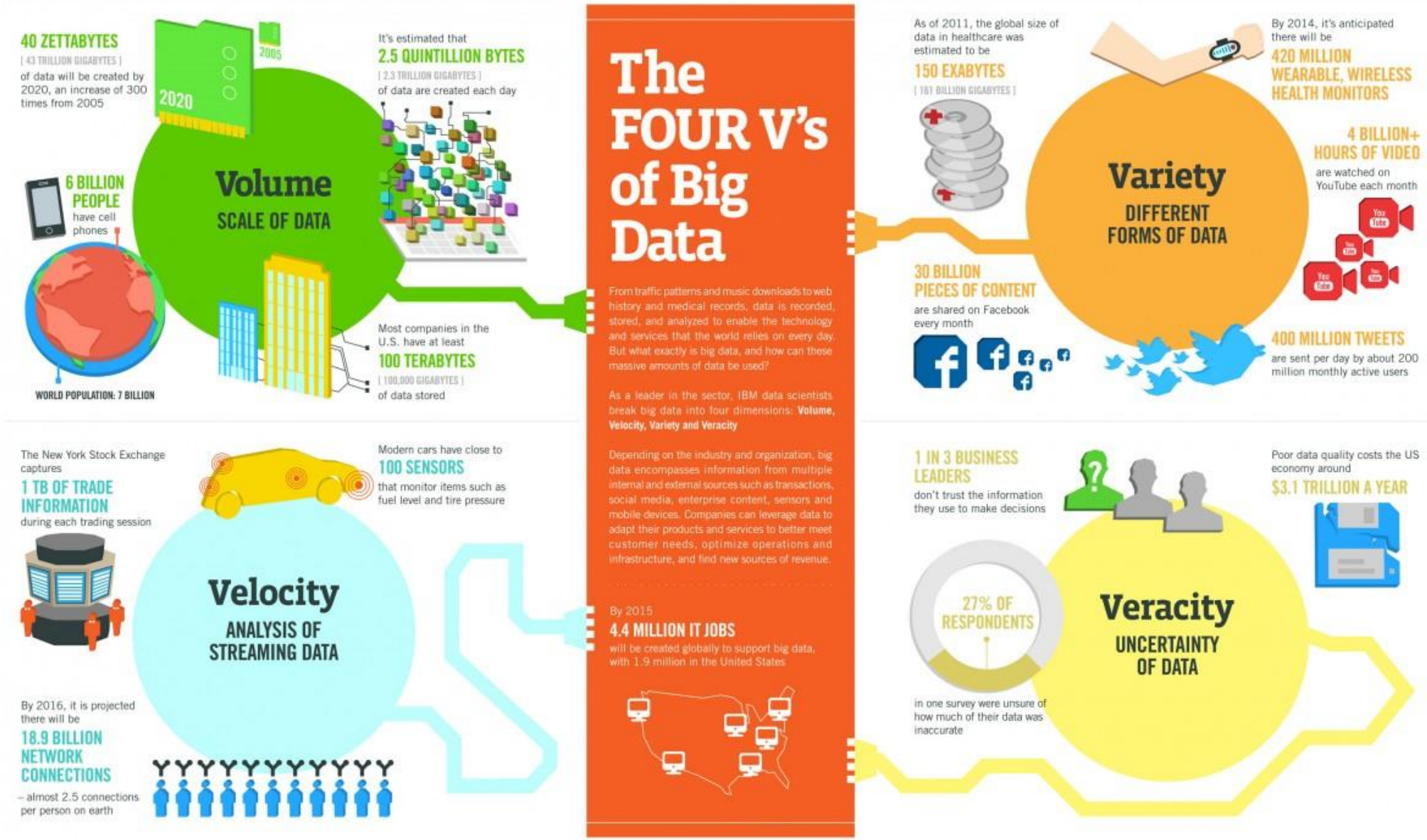
- Wprowadzenie w domenę
  - Założenia projektowe
  - Inteligentne liczniki
- Dlaczego chmura?
- Azure jako platforma dla dużych zbiorów danych
  - Ładowanie i przechowywanie danych
  - Przetwarzanie danych pomiarowych
- Q&A

FP Data Solutions

Założenia projektowe

- Dane przechowywane są w lokalnych DB klienta
- Dzienna porcja danych 15-20 GB (150-200 GB)
- Mamy ograniczony czas na przetworzenie danych
- Dane pobieramy każdego dnia o godzinie 6:00
- Konieczność sięgnięcia do danych historycznych – pesymistycznie 225 GB – 300 GB (2,25 TB – 3TB)

FP Data Solutions



FP Data Solutions

Czy możemy tutaj mówić o Big Data ?

- Wolumen 225 GB – 300 GB (2,25 TB – 3TB)
- Do 96 – 144 pomiarów dziennie na jeden licznik
- Dane pochodzące z różnych źródeł mają inną strukturę
- Obsługa danych niepoprawnych
- Możliwość uruchomienia z danymi zmodyfikowanymi

FP Data Solutions

Zaawansowana infrastruktura pomiarowa

Źródło: esdnews.com.au

FP Data Solutions

Zaawansowana infrastruktura pomiarowa

- 400 tysięcy liczników (domyślnie 4 mln.)
- AMI (Automatic Meter Infrastructure)
  - Liczniki
  - Koncentratory
  - Stacje
- PLC (*Power Line Communication*)

FP Data Solutions

Inteligentne liczniki

- Częstotliwość wykonywanych pomiarów
- 11 kanałów z danymi:
  - Moc czynna Ap +/-
  - Moc bierna pojemnościowa Rc +/-
  - Moc bierna indukcyjna Ri +/-
  - Natężenie prądu I
  - Napięcie prądu U

Źródło: Wikipedia

FP Data Solutions

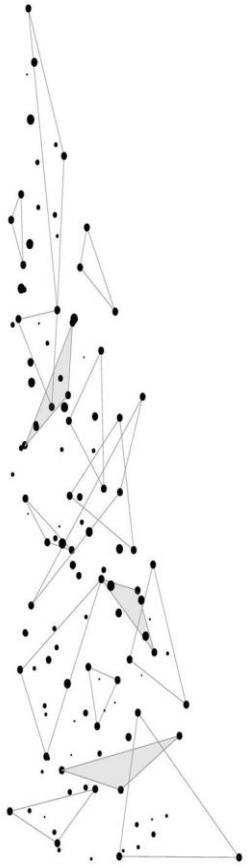
Analiza danych pomiarowych

Źródło: urbizedge.com/daas

FP Data Solutions

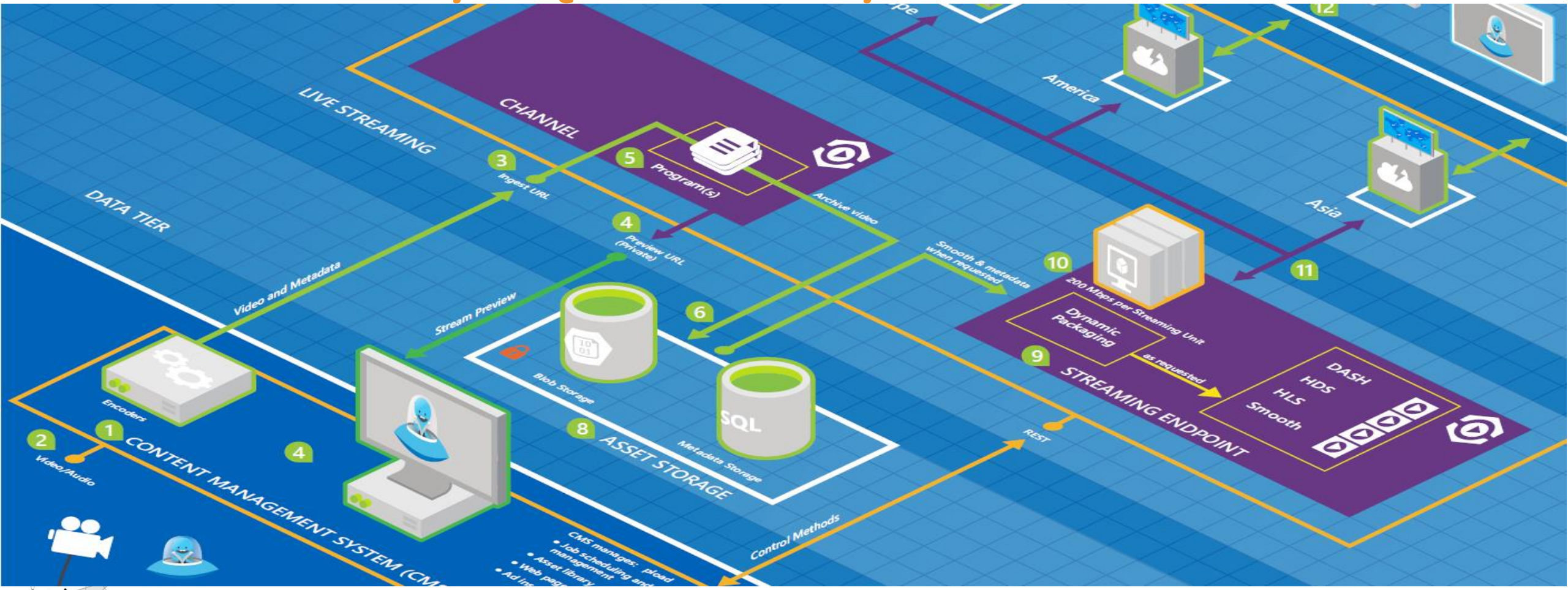
Wyzwania jakie zostały przed nami postawione

- Szacowanie braków w danych pomiarowych
  - Identyfikacja przerwy
  - Próba oszacowania brakujących pomiarów
- Wykrywanie anomalii
  - Nielegalny pobór energii elektrycznej
- Kalkulacja różnego rodzaju wskaźników i współczynników



FP Data Solutions

Zabieramy się do roboty!

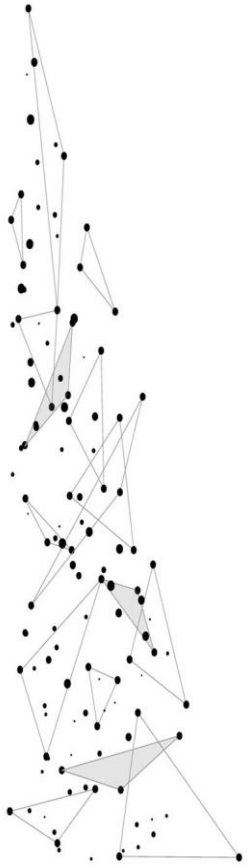


Źródło: urbizedge.com/daas

FP Data Solutions

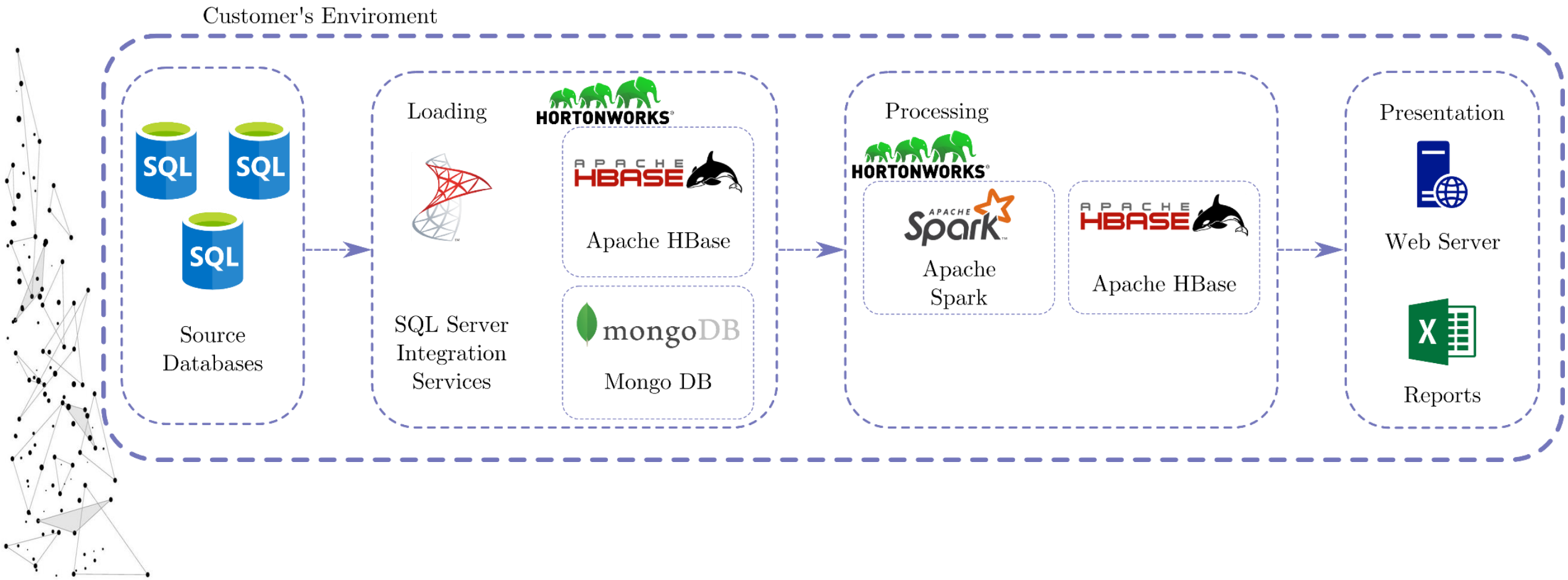
Co nam będzie potrzebne?

- Transfer wybranych danych do chmury
- Obszerny magazyn na dane (TB – PB)
- Automatyzacja i orkiestracja
- Przetwarzanie dużych zbiorów danych
  - Wybrane algorytmy do analizy danych pomiarowych
- Skalowalność



FP Data Solutions

Jakby to mogło wyglądać u klienta (on premise)



FP Data Solutions

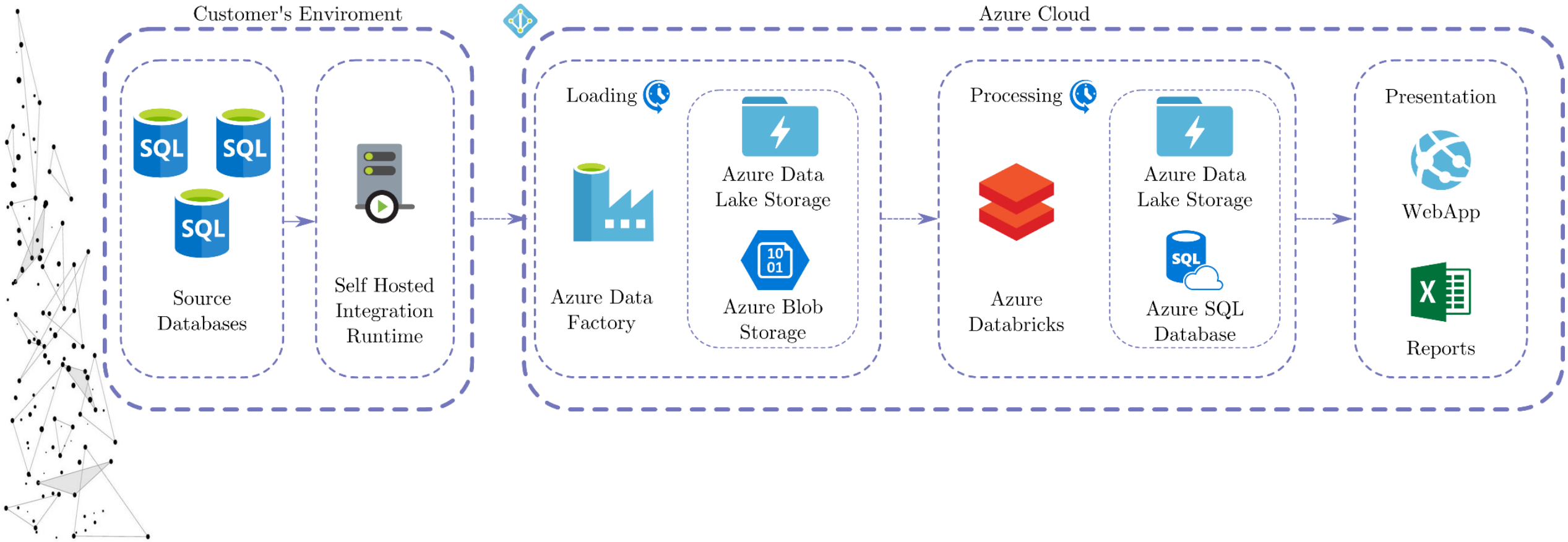
Dlaczego nie wybraliśmy tego podejścia?

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

computing © Jakub Jirsák - Fotolia.com

FP Data Solutions

Nasza architektura



FP Data Solutions

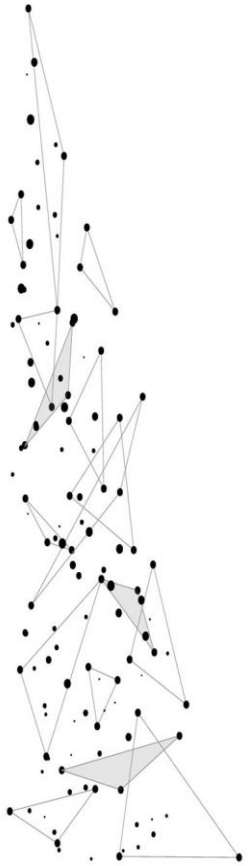
A co nam dał cloud?

- Nieograniczony storage
- Nieograniczona moc obliczeniowa
- Produktywność
- Dużo niższy próg wejścia
- Poczucie bezpieczeństwa (SLA)
- Koszty

FP Data Solutions

Co przeważyło nad cloudem

- Elastyczność
- Wysoka skalowalność i wydajność
- Niezawodność
- Optymalizacja kosztów
  - Płacimy za faktyczny czas użycia
- Bezpieczeństwo danych
  - Replikacja danych



amazon web services

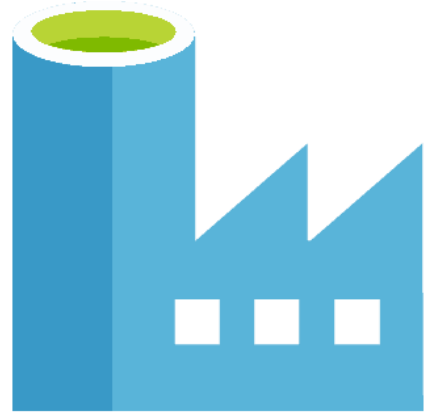
Azure

Google Cloud Platform

FP Data Solutions

Azure Data Factory

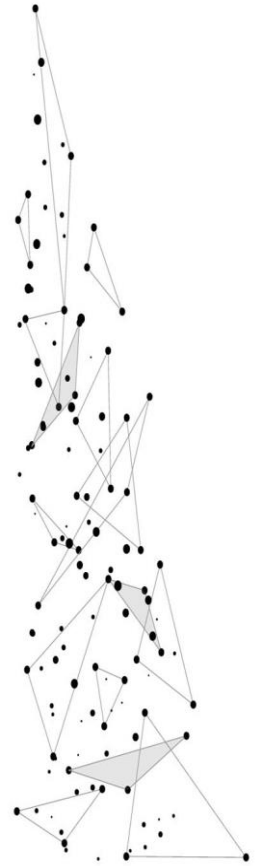
- Wysoka efektywność (Data Factory UI)
- Ponad 80 connectorów (ADLS, DataBricks)
- Umożliwia pełną automatyzację
- Integracja z środowiskami on premise
- Orkiestracja procesów w obrębie chmury
- Pozwala na wygodne przenoszenie i transformację danych
- Reagowanie na błędy



FP Data Solutions

Azure Data Lake Storage

- Wysoce wydajny i zoptymalizowany magazyn pod kątem dużego wolumenu danych
- Automatyczna replikacja danych
- Skalowalność
- Partycjonowanie danych
- Doskonała integracja z HDFS
- Teoretycznie nieograniczony magazyn na dane



FP Data Solutions

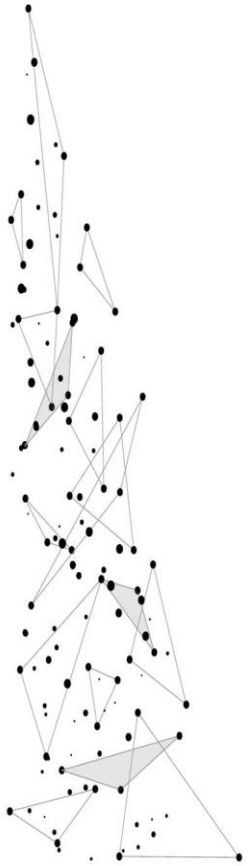
Azure Blob Storage

- Prosty i skalowalny magazyn
- Dane bez określonej struktury
- Automatyczna replikacja
- Integracja z projektowym CI (Azure CLI)

FP Data Solutions

Azure Databricks

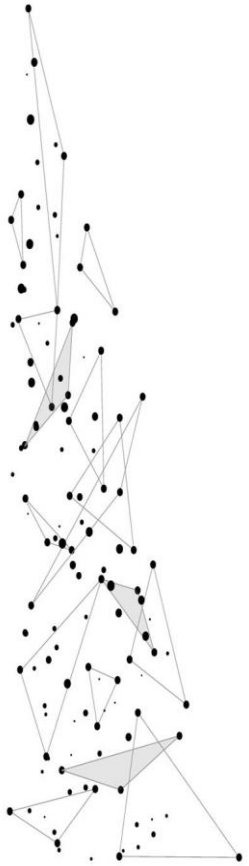
- Zoptymalizowane środowisko Apache Spark (CaaS)
- Automatyczne skalowanie
- Prosta integracja z innymi usługami PaaS'owych
- Pozwala na analizę w czasie rzeczywistym
- Wbudowane algorytmy uczenia maszynowego
- Wsparcie wielu języków programowania



FP Data Solutions

Azure SQL Server Database

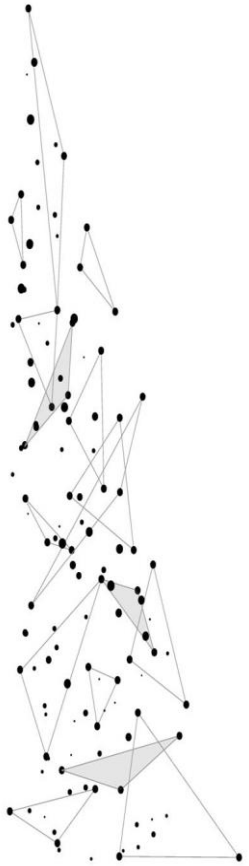
- Szybki dostęp do wybranych danych
- Inteligentne przetwarzanie zapytań
- Inteligentne dostrajanie wydajności
- Monitorowanie
- Skalowalność

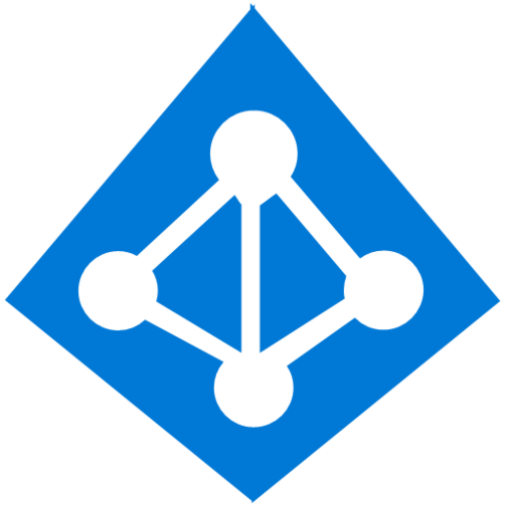


FP Data Solutions

Azure Active Directory

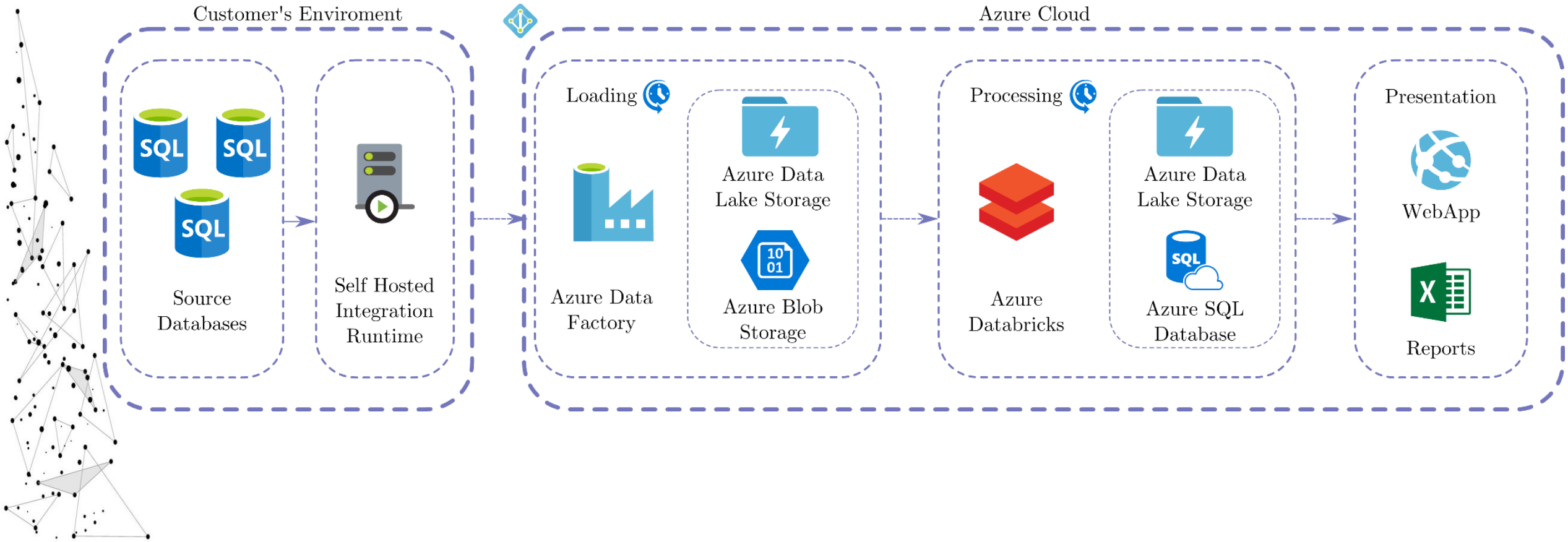
- Bezproblemowy i wysoce bezpieczny dostęp
- Ochrona tożsamości użytkowników i developerów
- Prosta i wygodna kontrola nad rolami
- Ograniczenie dostępu do wybranych usług





FP Data Solutions

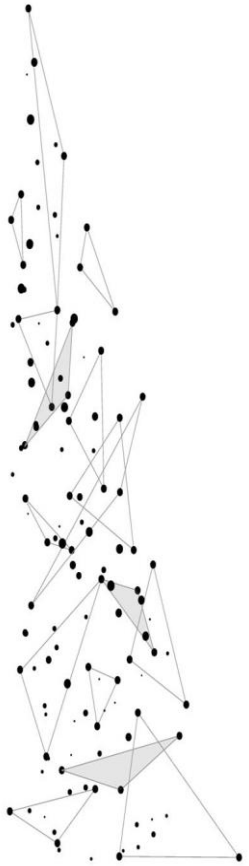
Nasza architektura



FP Data Solutions

Jakie napotkaliśmy problemy i co zrobiliśmy źle?

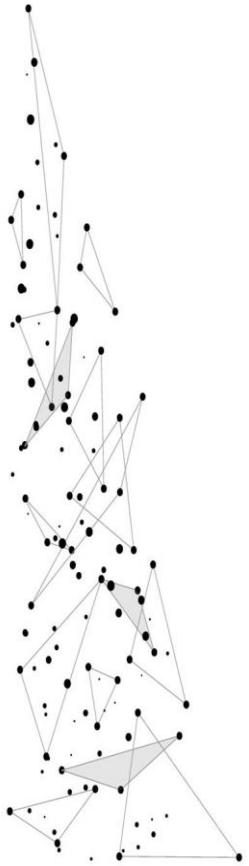
- Czy Azure był najlepszym wyborem?
- PySpark zamiast natywnego Spark'a
- Zbyt mały budżet na niektóre usługi
  - Łatwo przetrwonić pieniądze



FP Data Solutions

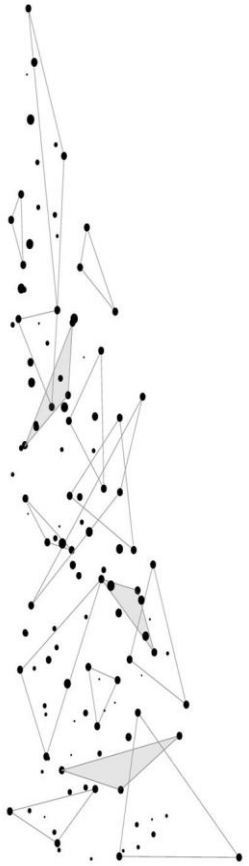
Wnioski

- Cloud stanowi bardzo dobre środowisko dla projektów Big Data o ruchomej skali problemu
- Chmura pozwala na o wiele szybszy start
- Krótkoterminowe projekty lepiej odnajdują się w chmurze
- Storage jest tani, usługi związane z przetwarzaniem danych są drogie
- Nie zawsze musimy brać za wszystko pełną odpowiedzialność



FP Data Solutions

Pytanie



Za co jesteście gotowi wziąć odpowiedzialność?

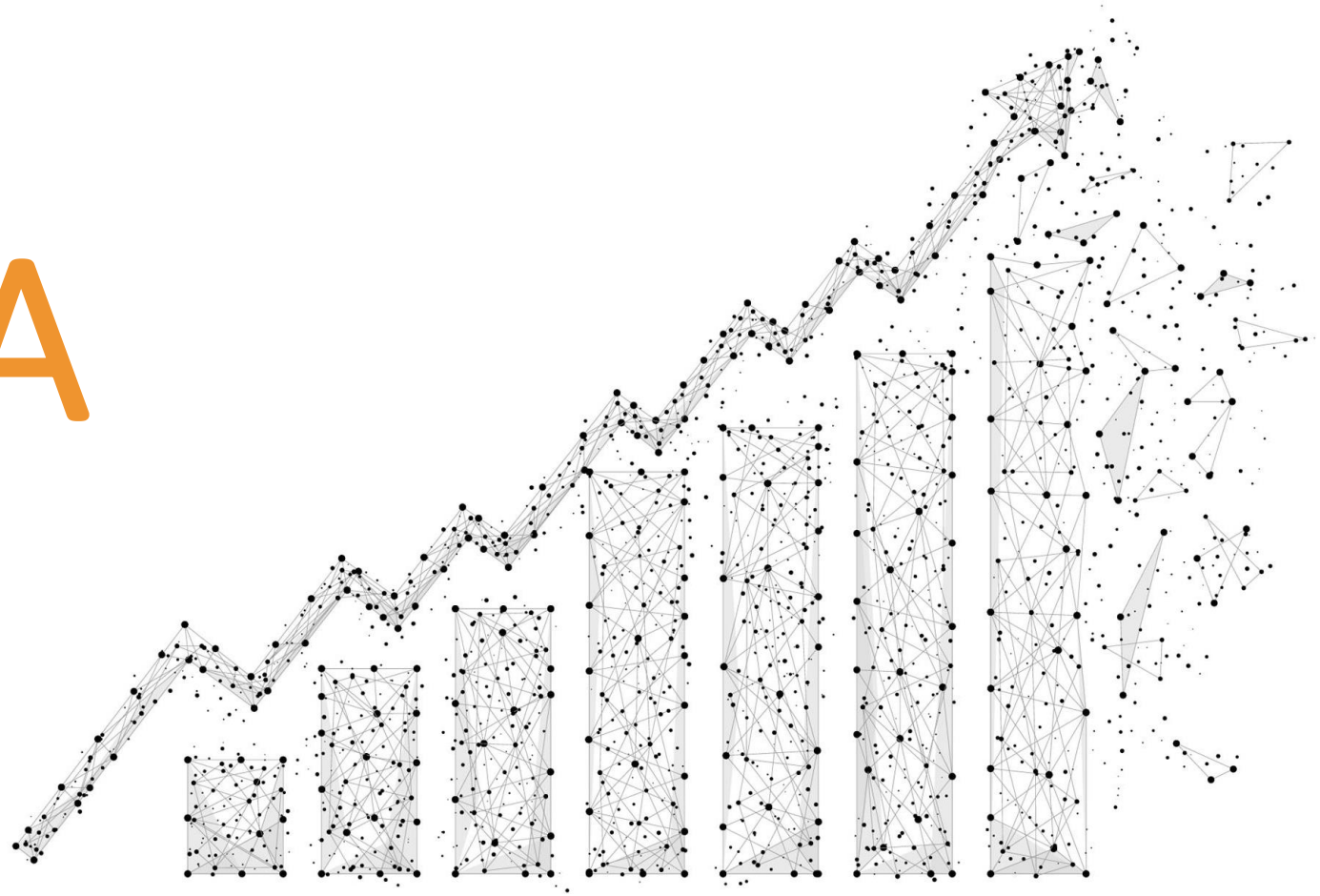
FP Data Solutions

Resources

- https://docs.microsoft.com/pl-pl/azure/data-factory/
- https://docs.microsoft.com/pl-pl/azure/storage/blobs/data-lake-storage-introduction
- https://azure.microsoft.com/pl-pl/services/databricks/
- https://docs.databricks.com/

FP Data Solutions

Q & A



www.future-processing.com

FP Data Solutions

THANK YOU!

www.future-processing.com