



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

Kamil Dworak



O mnie

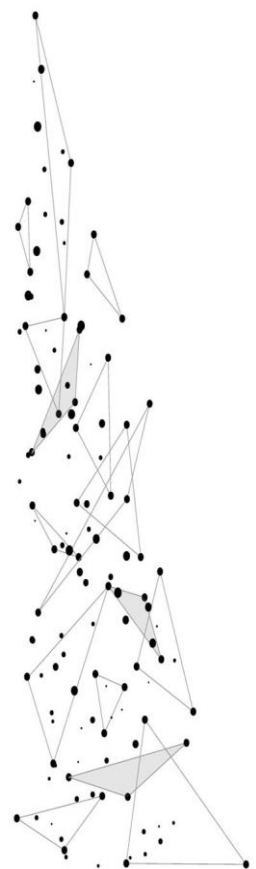
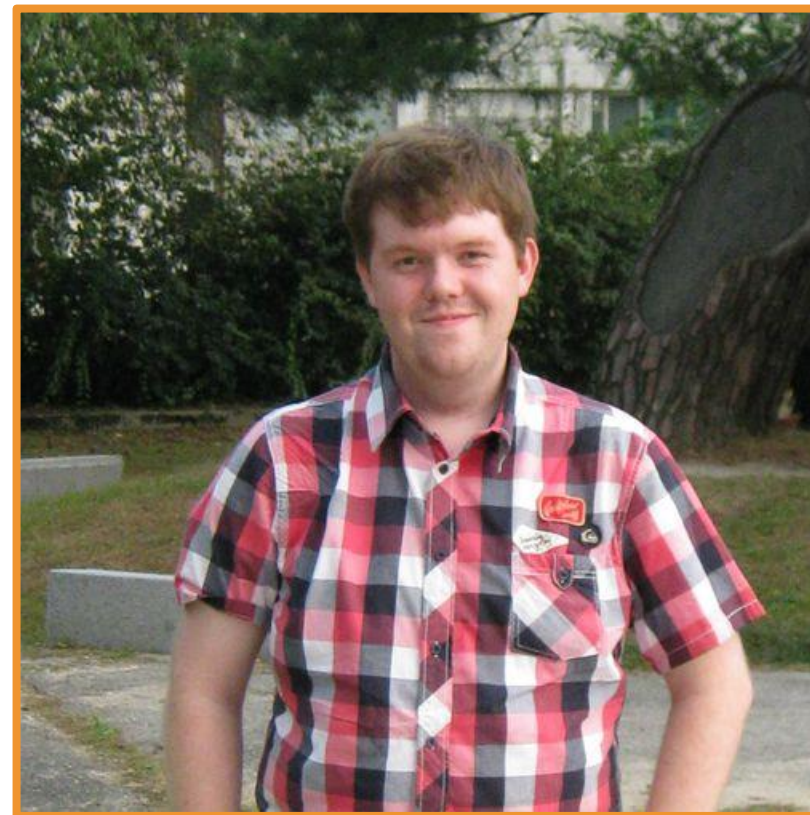
Kamil Dworak

 **FP Data Solutions**

 **UNIwersYTET ŚLĄSKI**
W KATOWICACH

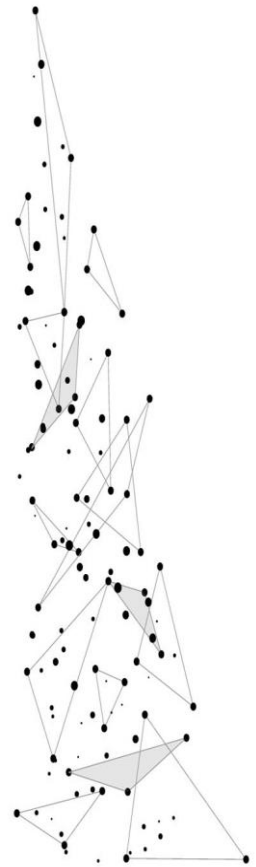
 <https://github.com/devkam/Events/>

 <http://bujacwoblokach.pl/>



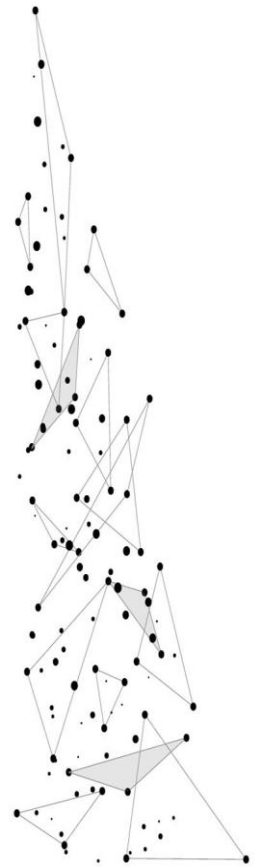
Agenda

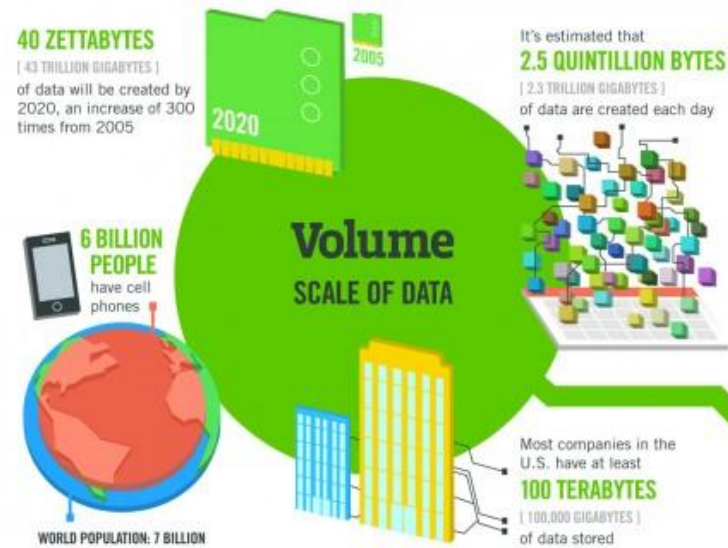
- Wprowadzenie w domenę
 - Założenia projektowe
 - Inteligentne liczniki
- Dlaczego chmura?
- Azure jako platforma dla dużych zbiorów danych
 - Ładowanie i przechowywanie danych
 - Przetwarzanie danych pomiarowych
- Q&A



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)





The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

Variety
DIFFERENT FORMS OF DATA

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to **100 SENSORS**
that monitor items such as fuel level and tire pressure

Velocity
ANALYSIS OF STREAMING DATA

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
— almost 2.5 connections per person on earth



1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

Poor data quality costs the US economy around
\$3.1 TRILLION A YEAR

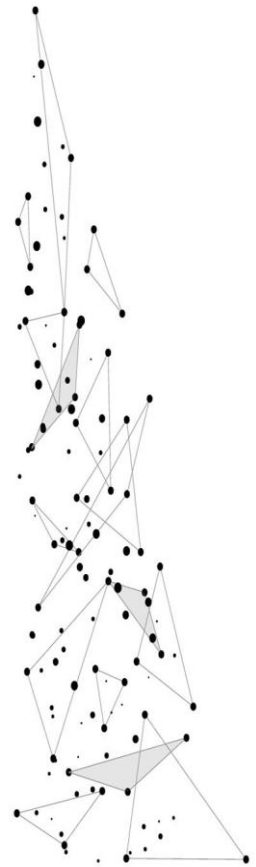
Veracity
UNCERTAINTY OF DATA

27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

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



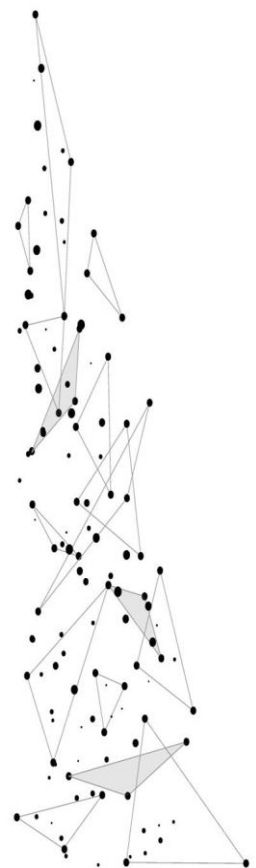
Zaawansowana infrastruktura pomiarowa



Źródło: esdnews.com.au

Zaawansowana infrastruktura pomiarowa

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



Inteligentne liczniki

- Częstotliwość wykonywanych pomiarów
- 11 kanałów z danymi:
 - Moc czynna A_p +/-
 - Moc bierna pojemnościowa R_c +/-
 - Moc bierna indukcyjna R_i +/-
 - Natężenie prądu I
 - Napięcie prądu U



Źródło: [Wikipedia](https://en.wikipedia.org/wiki/Meter)

Analiza danych pomiarowych

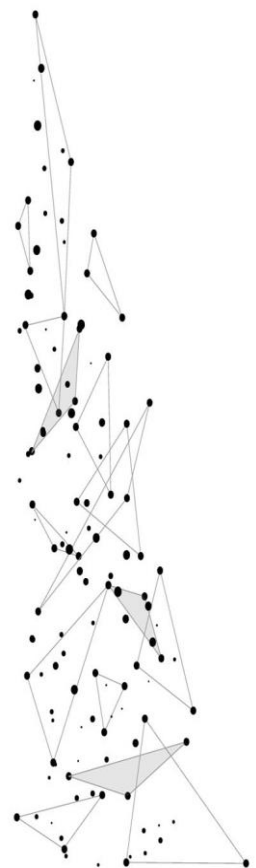


Źródło: urbizedge.com/daas



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

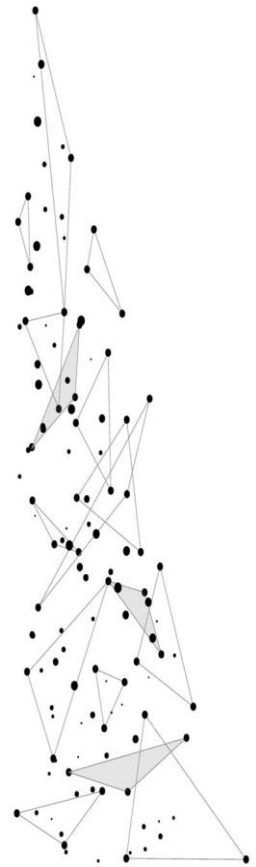


Zabieramy się do roboty!

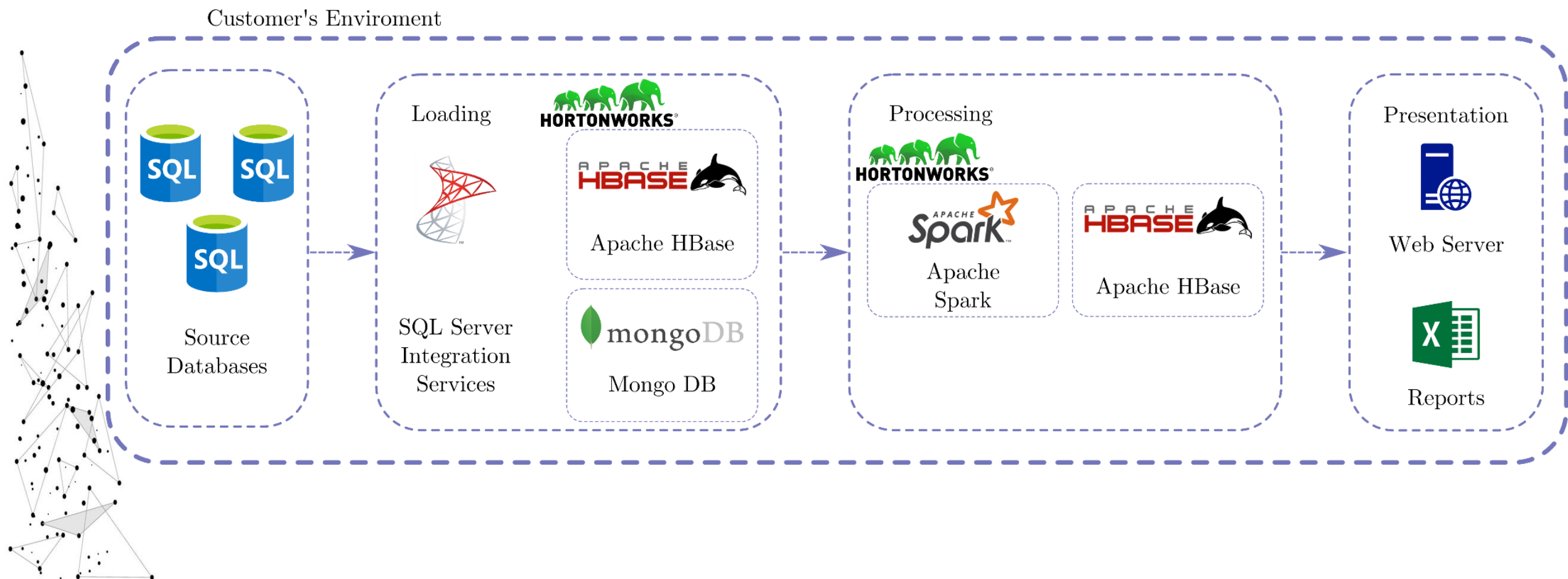


Co nam będzie potrzebne?

- 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

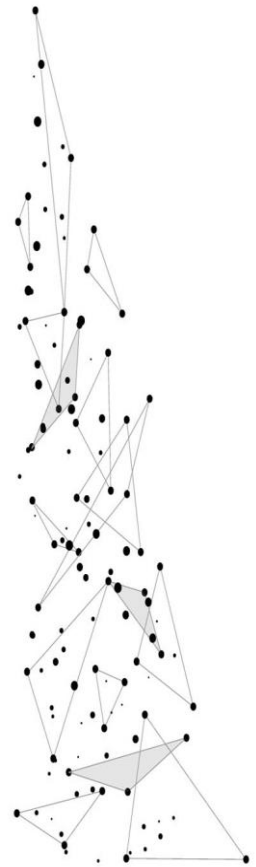


Jakby to mogło wyglądać u klienta (on premise)



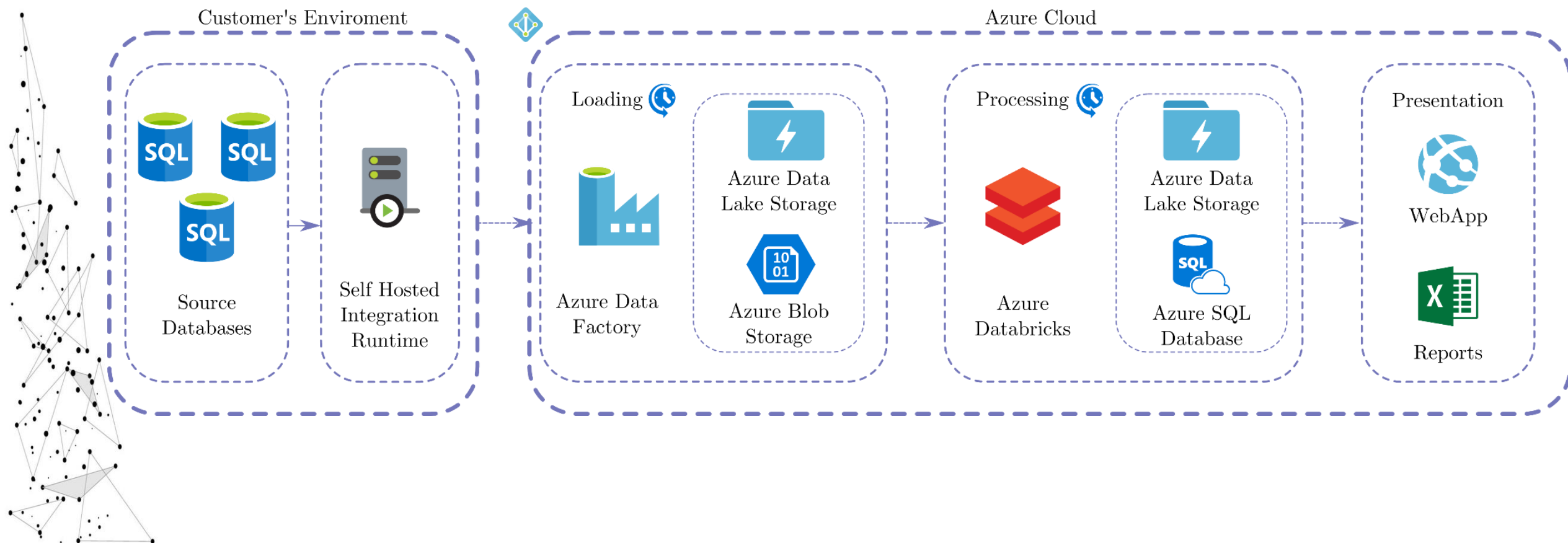
Dlaczego nie wybraliśmy tego podejścia?

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



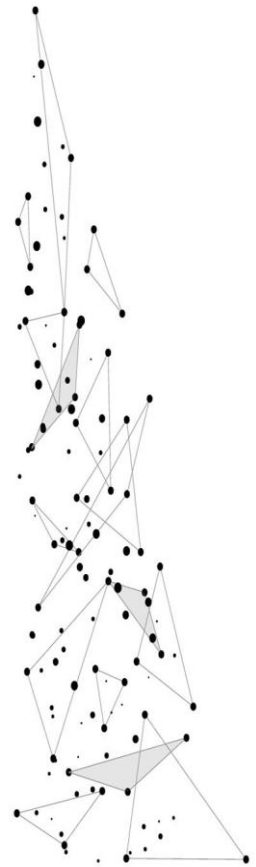


Nasza architektura



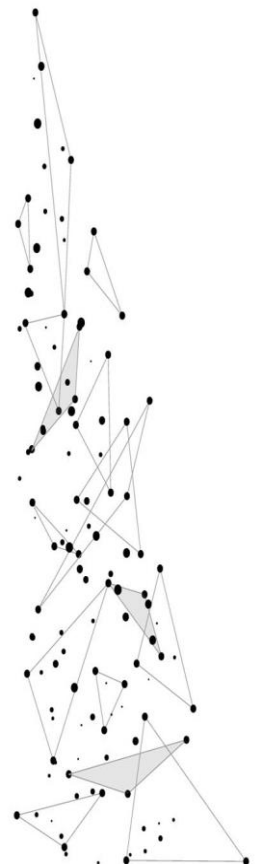
A co nam dał cloud?

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



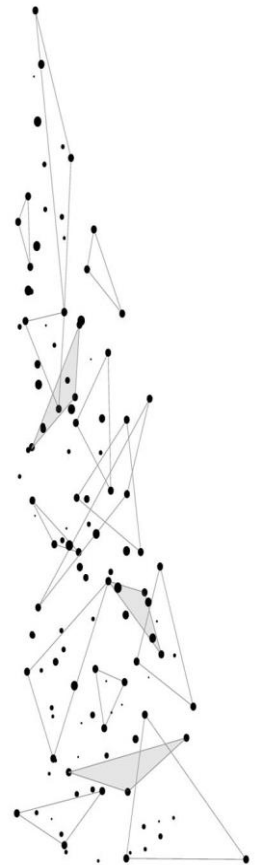
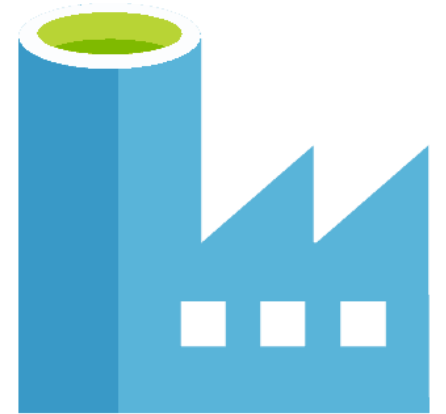
Co przeważało nad cloudem

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



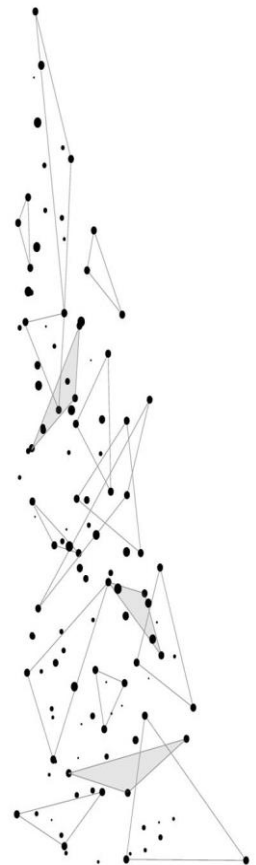
Azure Data Factory

- 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



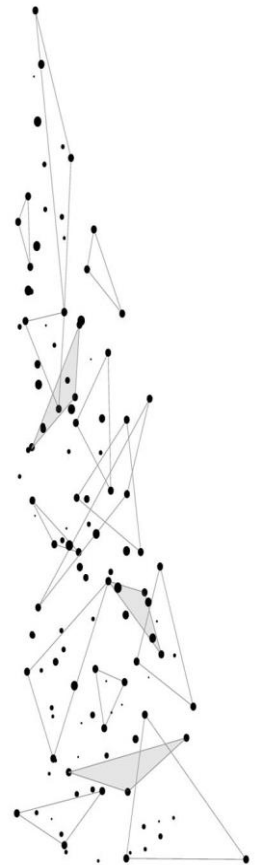
Azure Data Lake Storage

- 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



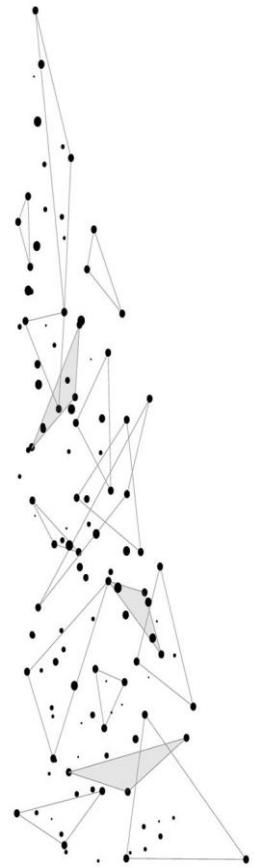
Azure Blob Storage

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



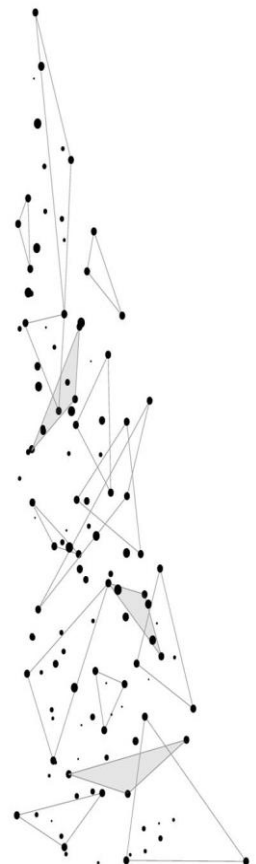
Azure Databricks

- 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



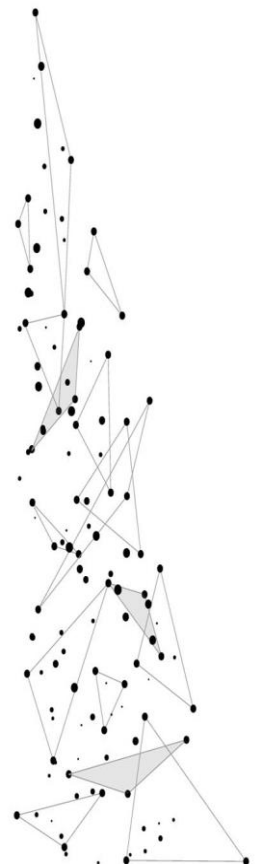
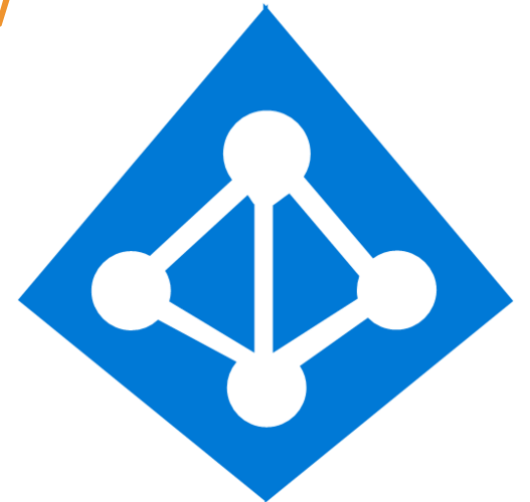
Azure SQL Server Database

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

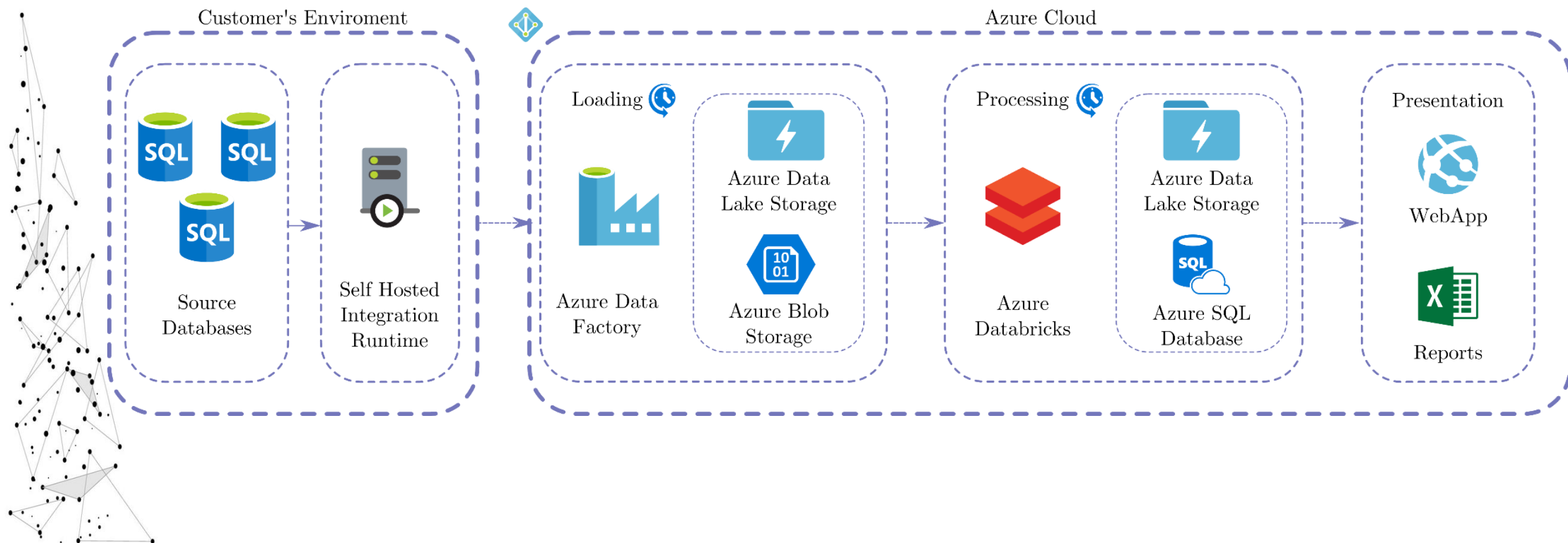


Azure Active Directory

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

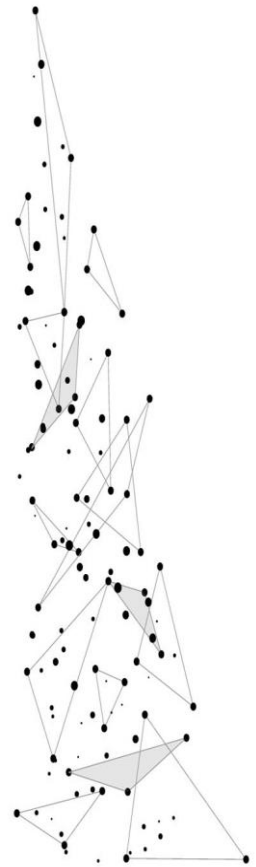


Nasza architektura



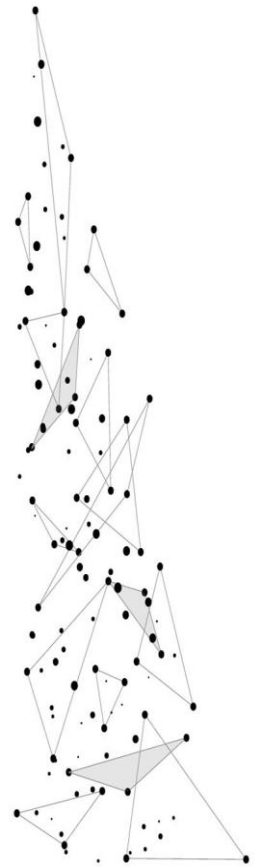
Jakie napotkaliśmy problemy i co zrobiliśmy źle?

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



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



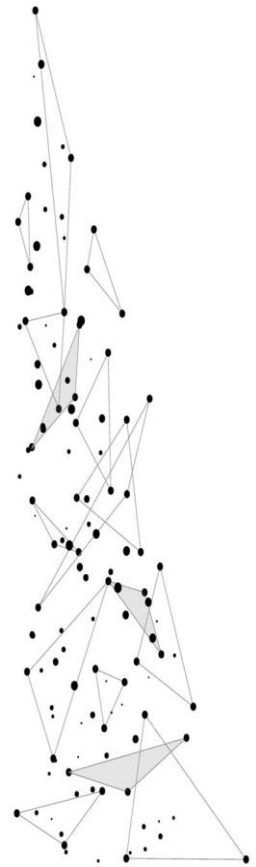
Pytanie

Za co jesteście gotowi wziąć
odpowiedzialność?



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/>



Q & A

THANK YOU!