

Analiza danych okiem praktyka

Jak szybko zacząć i szczęśliwie skończyć?

Michał Żyliński

michal.zylinski@microsoft.com

Microsoft

O mnie

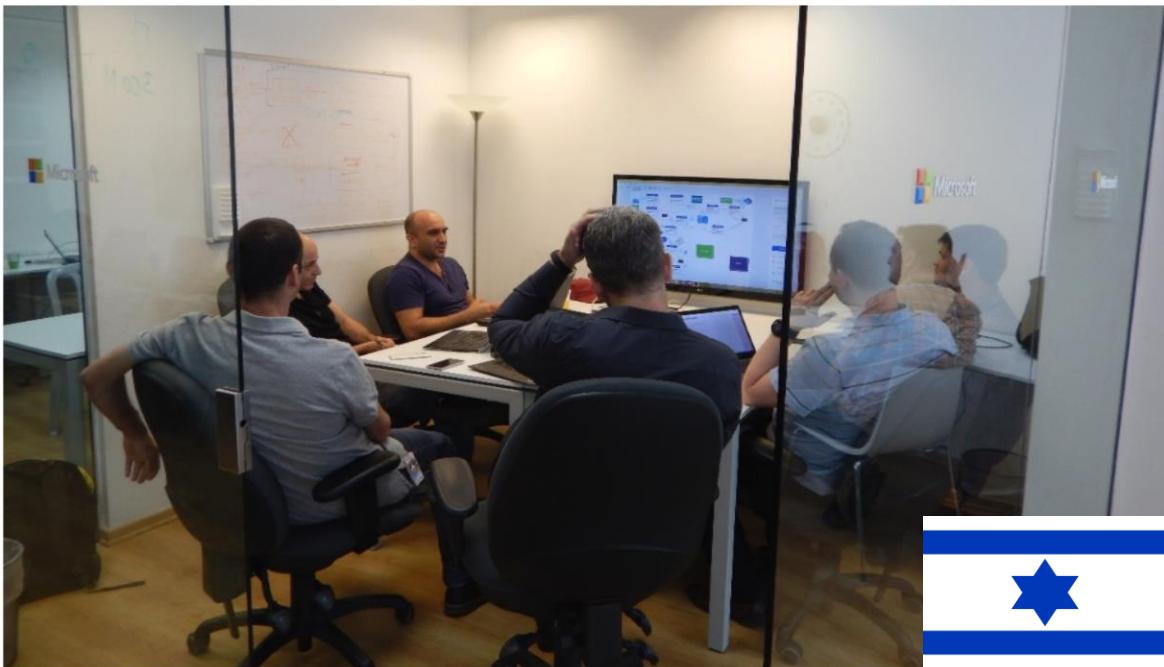
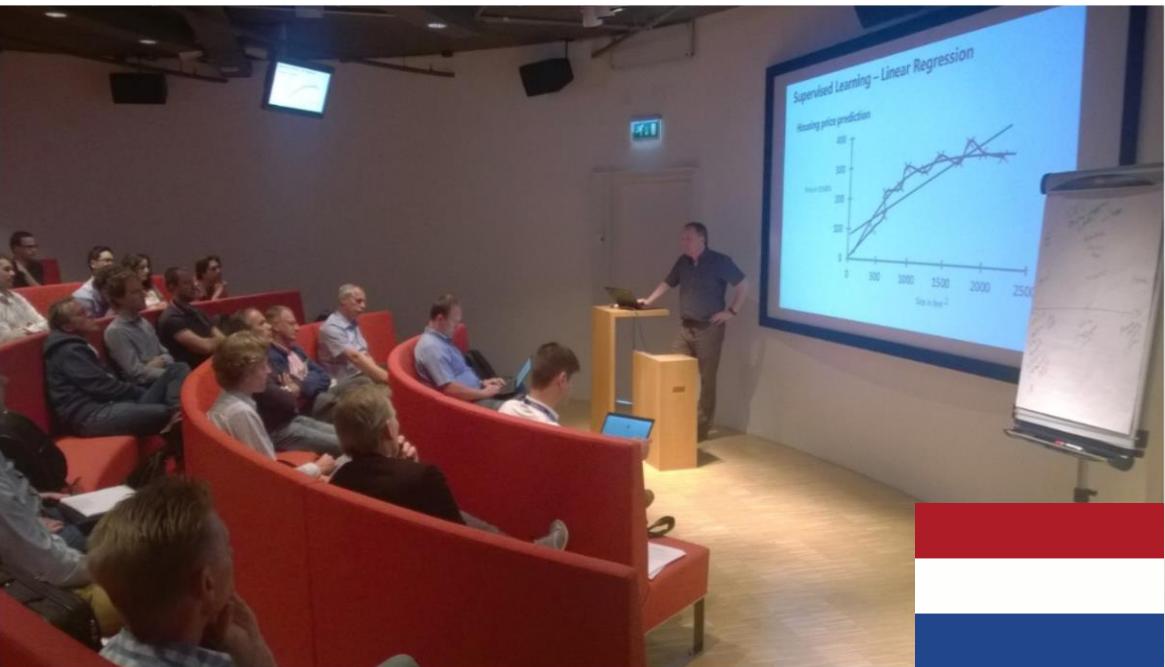
- 9 lat w Microsoft
- „Ostatnio” w zespole Advanced Analytics Global Blackbelts
- Deweloper, nie DBA
- Debiutant na SQLDay ☺

Cele sesji

- Omówienie typowych scenariuszy analitycznych
- Wskazanie dobrych praktyk i często popełnianych błędów
- Określenie kryteriów sukcesu projektowego

Cele sesji

- Wymiana doświadczeń
- Inspiracja
- Pomoc
- 7 klientów
- 7 reguł
- 0,5 ~~demo~~ żywego przykładu



Dlaczego?

- Otwartość
- Technologia
- Ludzie
- Integracja
- Chmura



Więcej: <http://www.zdnet.com/article/microsofts-r-strategy/>

Forrester Wave: Big data Hadoop Cloud Solutions, Q2 2016

Źródło: <https://azure.microsoft.com/en-us/blog/forrester-names-microsoft-azure-a-leader-in-big-data-hadoop-cloud-solutions/>

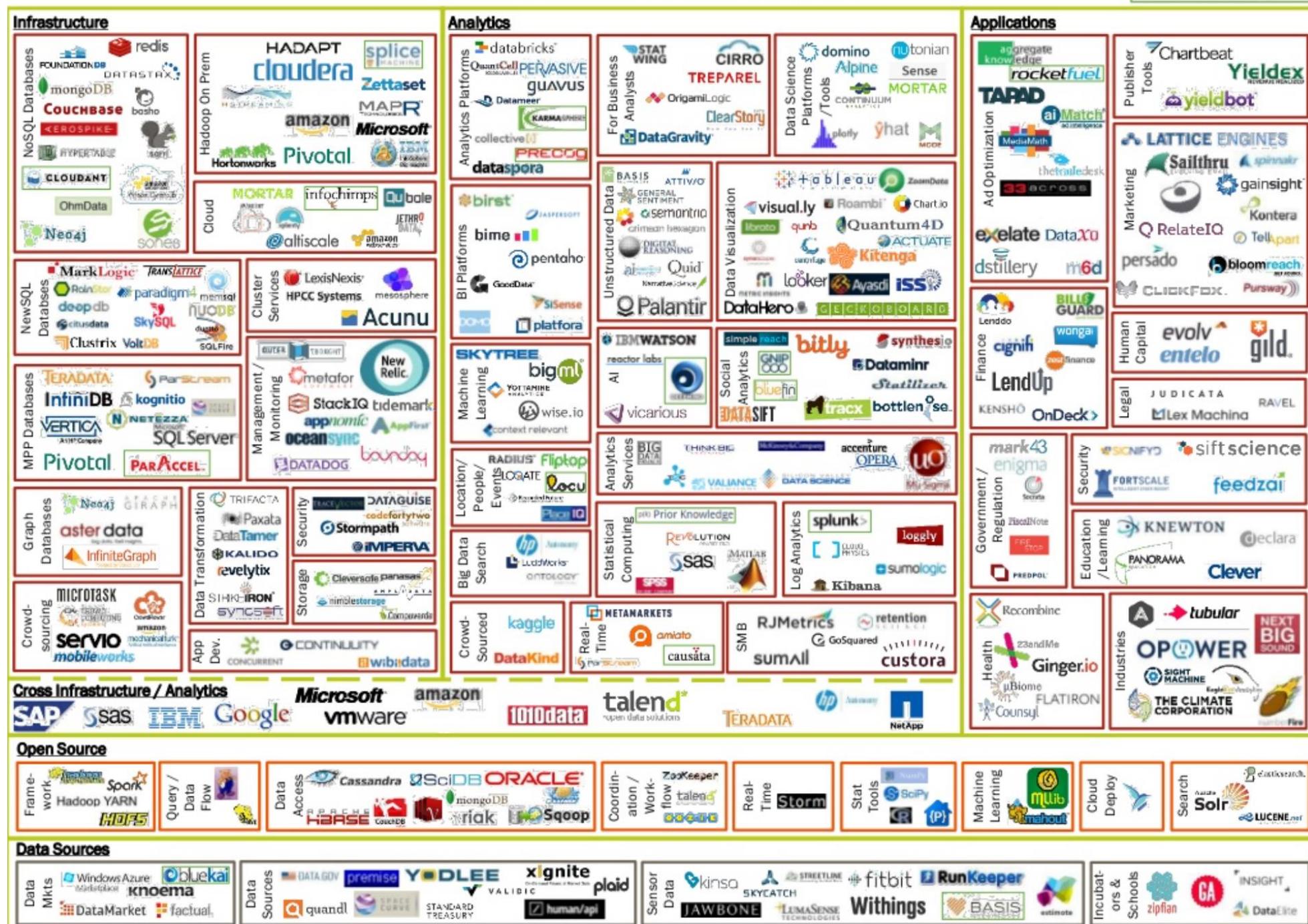
Kieruj się
prostota



<http://www.photos-public-domain.com/2011/09/11/simple/>

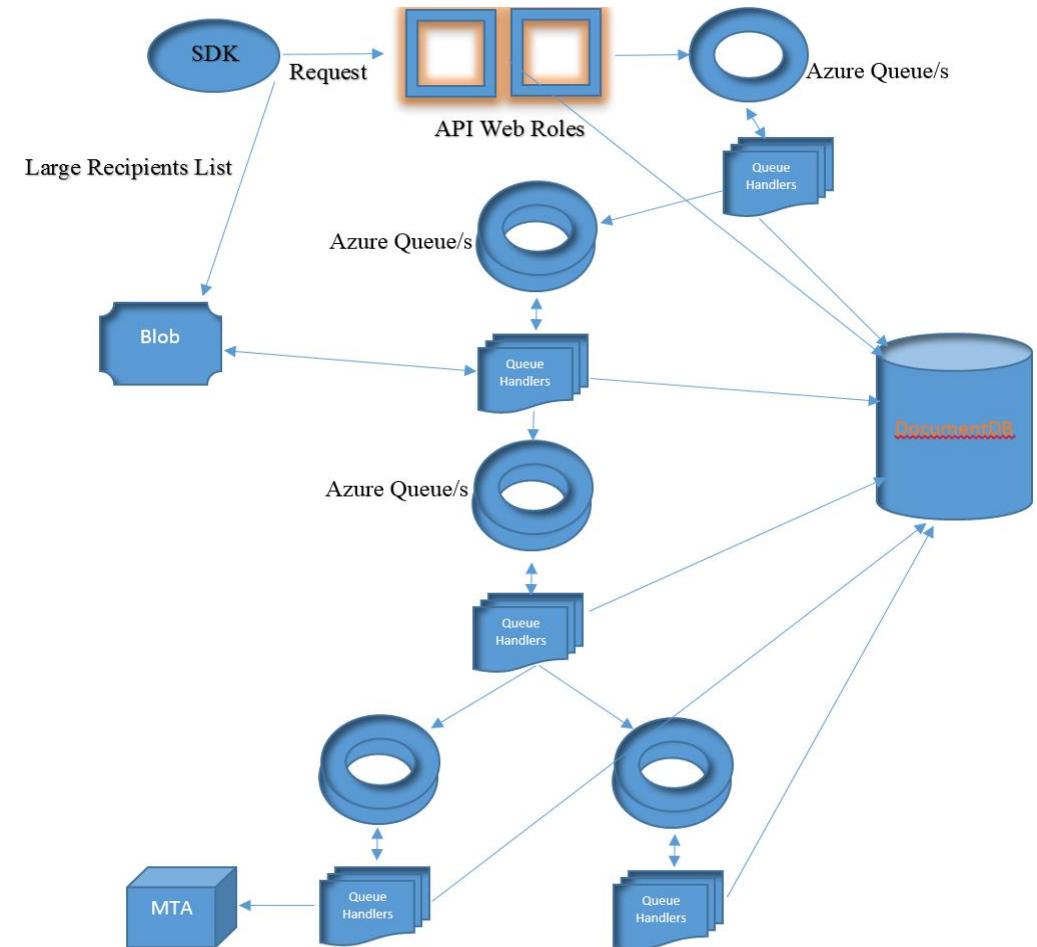
BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO



Przypadek #1: big data... lub nie

- Rozwiązanie działające już w chmurze, ale wymagające pilnej zmiany architektury ze względu na skalowalność i kończące się wsparcie dla SQL Federations
- MongoDB na dokładkę



Przypadek #1: Rezultaty

- Rozważano wiele opcji (DocumentDB, Hadoop, SQL Server), ale ostateczna decyzja była prosta- migracja do SQL DB Premium
 - Brak zmian w architekturze
 - Wykorzystanie dotychczasowych kompetencji zespołu
 - Wystarczająca przestrzeń na wzrost
 - Prostota!!!

Ufaj chmurze



<http://www.flickr.com/photos/mw7/>

Przypadek #2: PaaS kontra IaaS

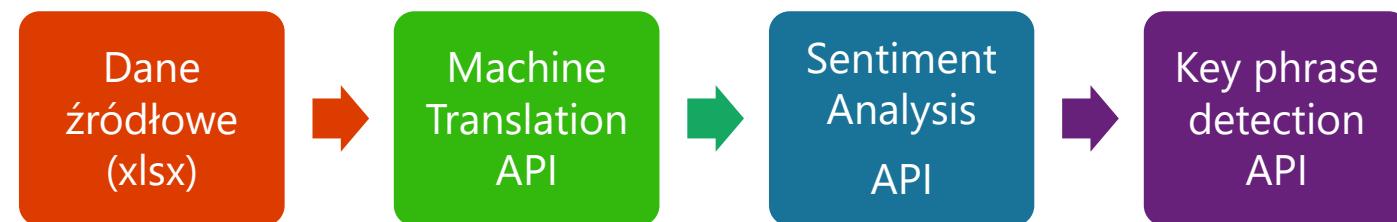
- Klient korzystający do tej pory z SQL Database podjął decyzję o zmianie w warstwie danych
- Wybrano HBase, chociaż zespół nie miał wcześniej doświadczeń z Hadoopem (do tej pory środowisko Microsoft)
- Specyfika wymagań wymusiła „ręczną” instalację dodatkowych bibliotek (np. silnika indeksującego). Mimo to zdecydowano się na usługę HDInsight (PaaS).

Przypadek #2: Rezultaty

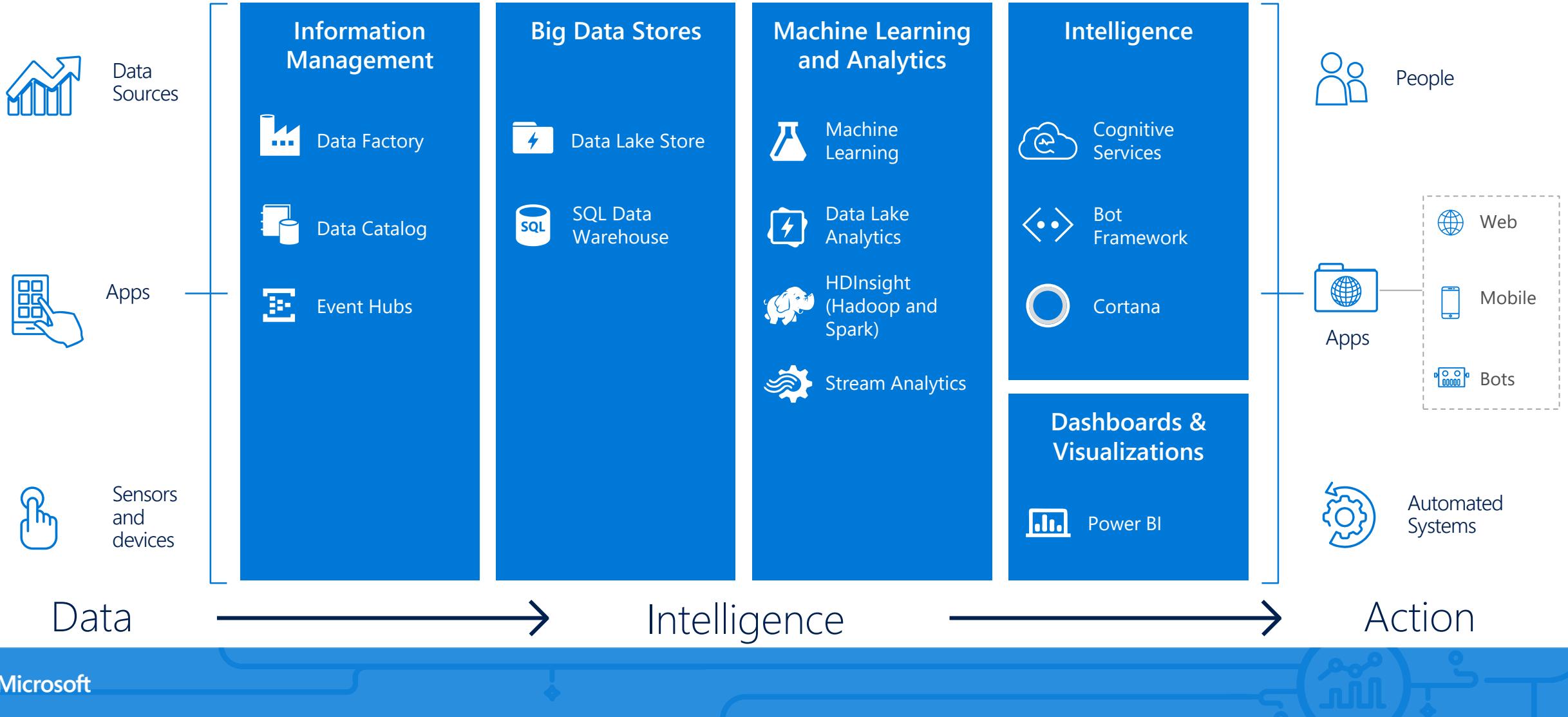
- Wdrożenie systemu produkcyjnego nastąpiło w ciągu 4 miesięcy od pierwszego spotkania architektonicznego
- Większość wyzwań została w tym czasie rozwiązana
- Niestety po starcie zaczęły pojawiać się dziwne zjawiska
- Dzięki modelowi PaaS zaangażowano bezpośrednio inżynierów z grupy produktowej
- Wniosek: przy wyborze technologii (również chmurowych) weź pod uwagę typ i zakres wsparcia producenta

Przypadek #3: Spróbuj zamiast narzekać

- Jak przeprowadzić analizę tekstu (sentymet)?
- Czy Cortana API da radę?
- Zalety:
 - Rozwiązanie niemalże z pudełka
 - Nie potrzebujemy wiedzy dziedzinowej
- Minusy:
 - Wsparcie tylko języka angielskiego
 - Brak elastyczności
 - Czarna skrzynka



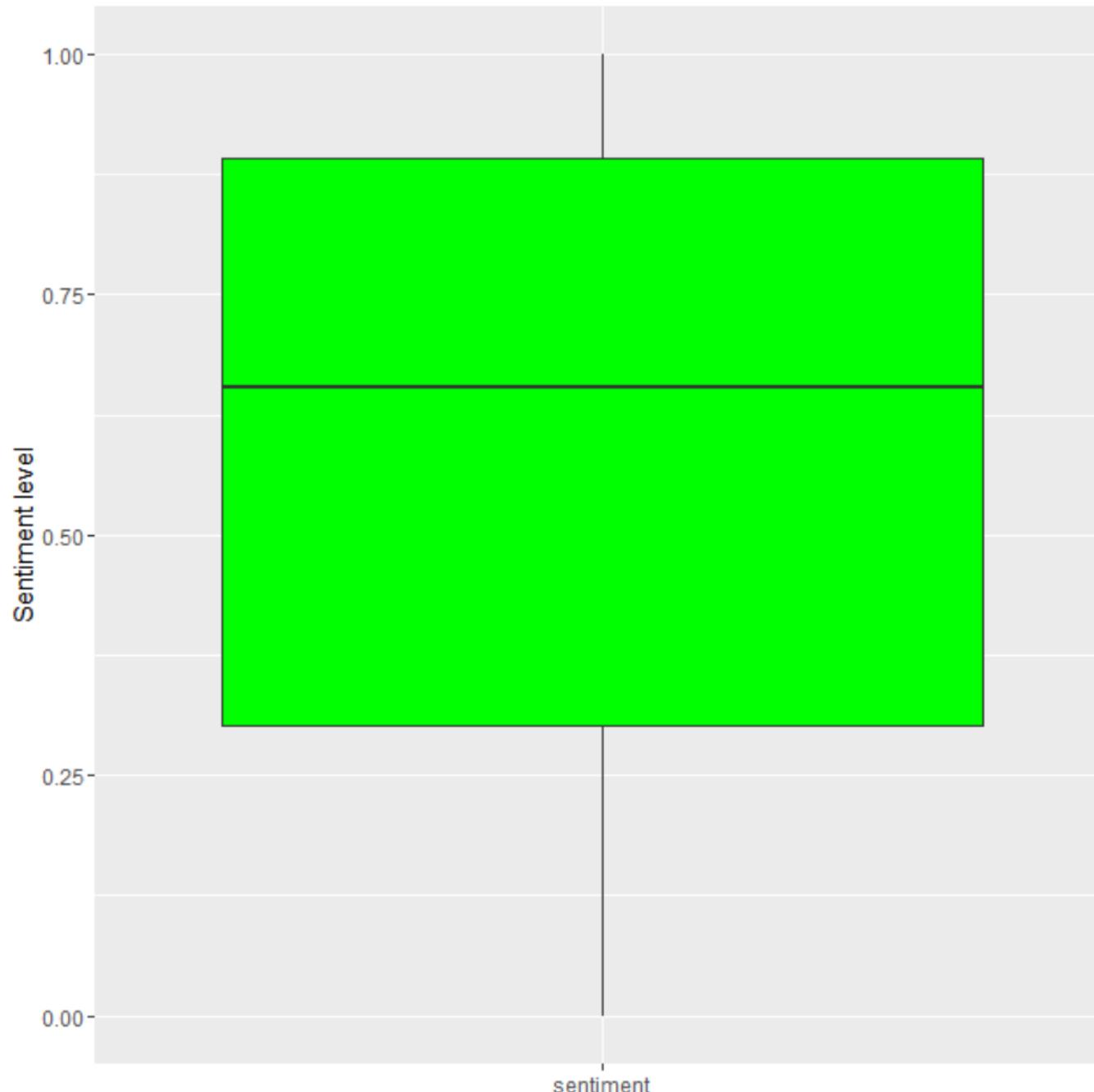
Cortana Intelligence Suite



Sprzedawaj historie



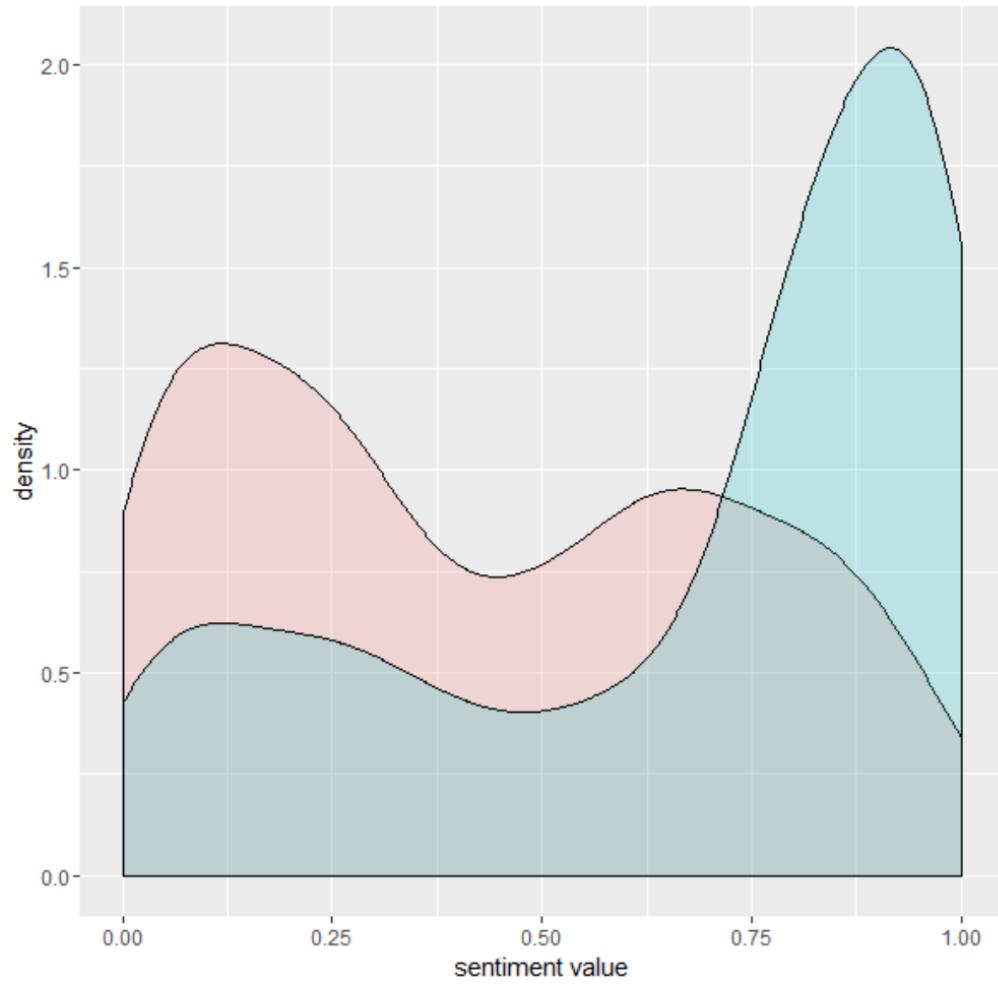
https://commons.wikimedia.org/wiki/File:Bedtime_story_-_Madeline.JPG



different corporate

employees

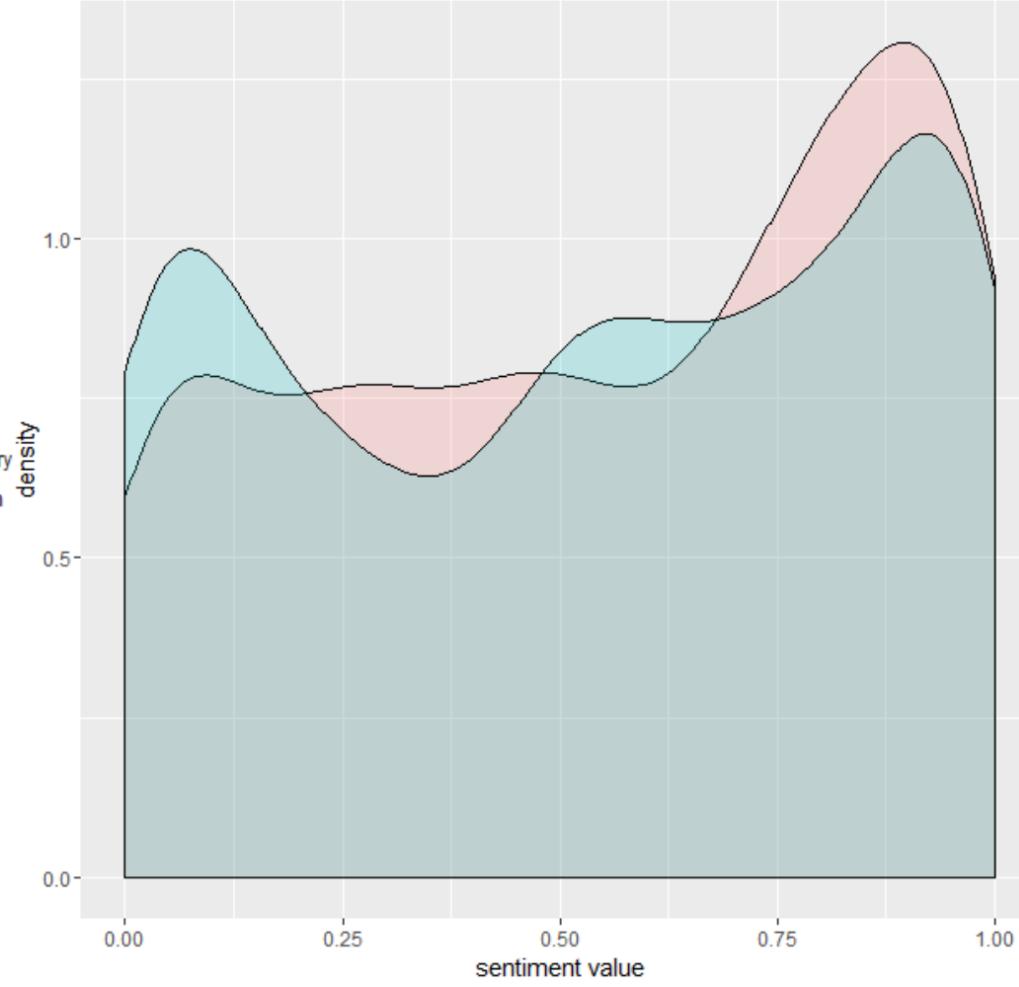
training
new process constant processes
real terms products sales high issues problems program thanks experience
processes growth income level head reduction salaries additional internal office
customers motivation business job career department increase financial number team services
units manager gosb wages day small years great time best lack quality
performance good people large payment day working salary lot changes huge client clients
lot changes huge client clients
service necessary
work
bank
corporate
customer



Topic

salary

team



Topic

future|changes

leadership|management

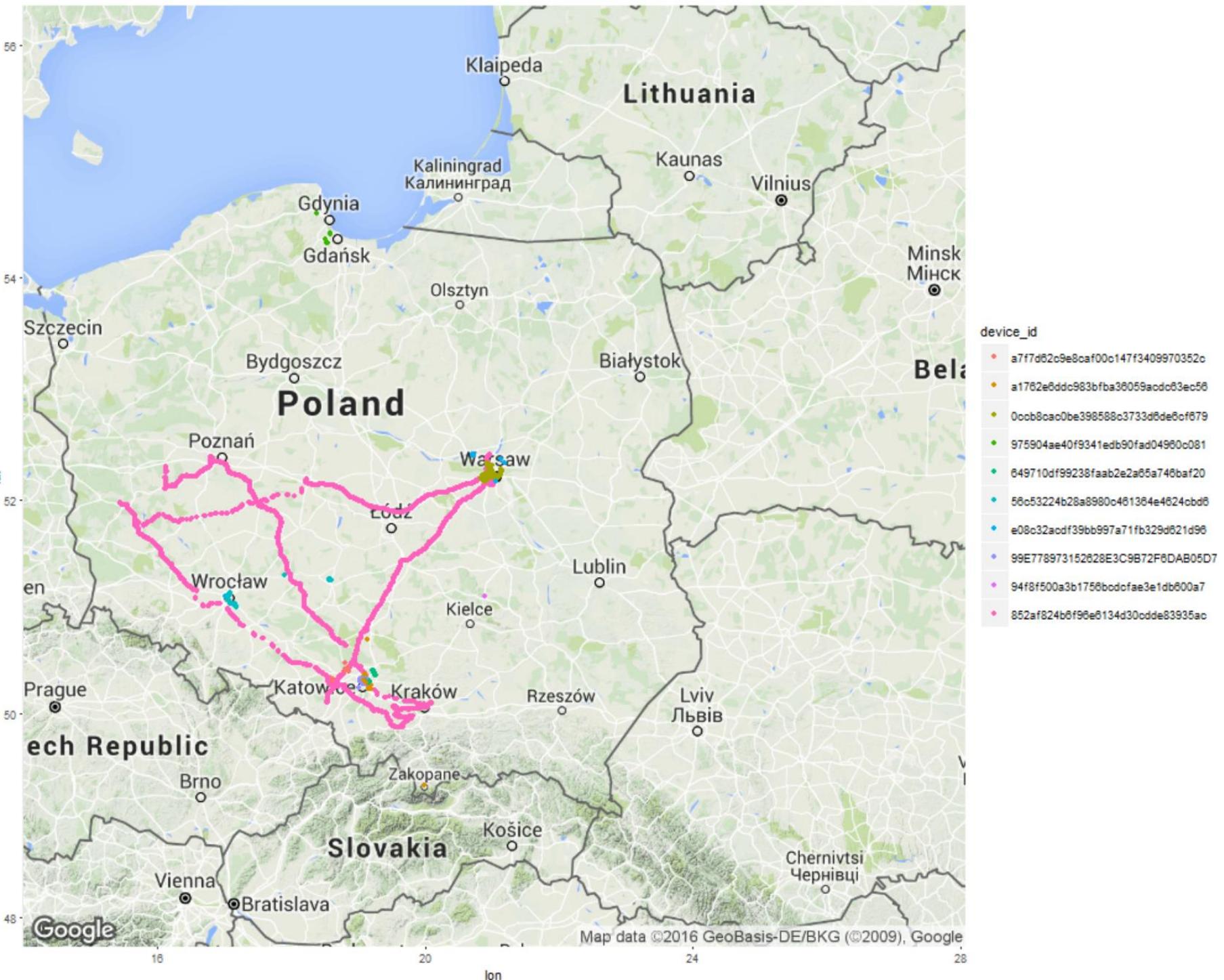
Inspiruj

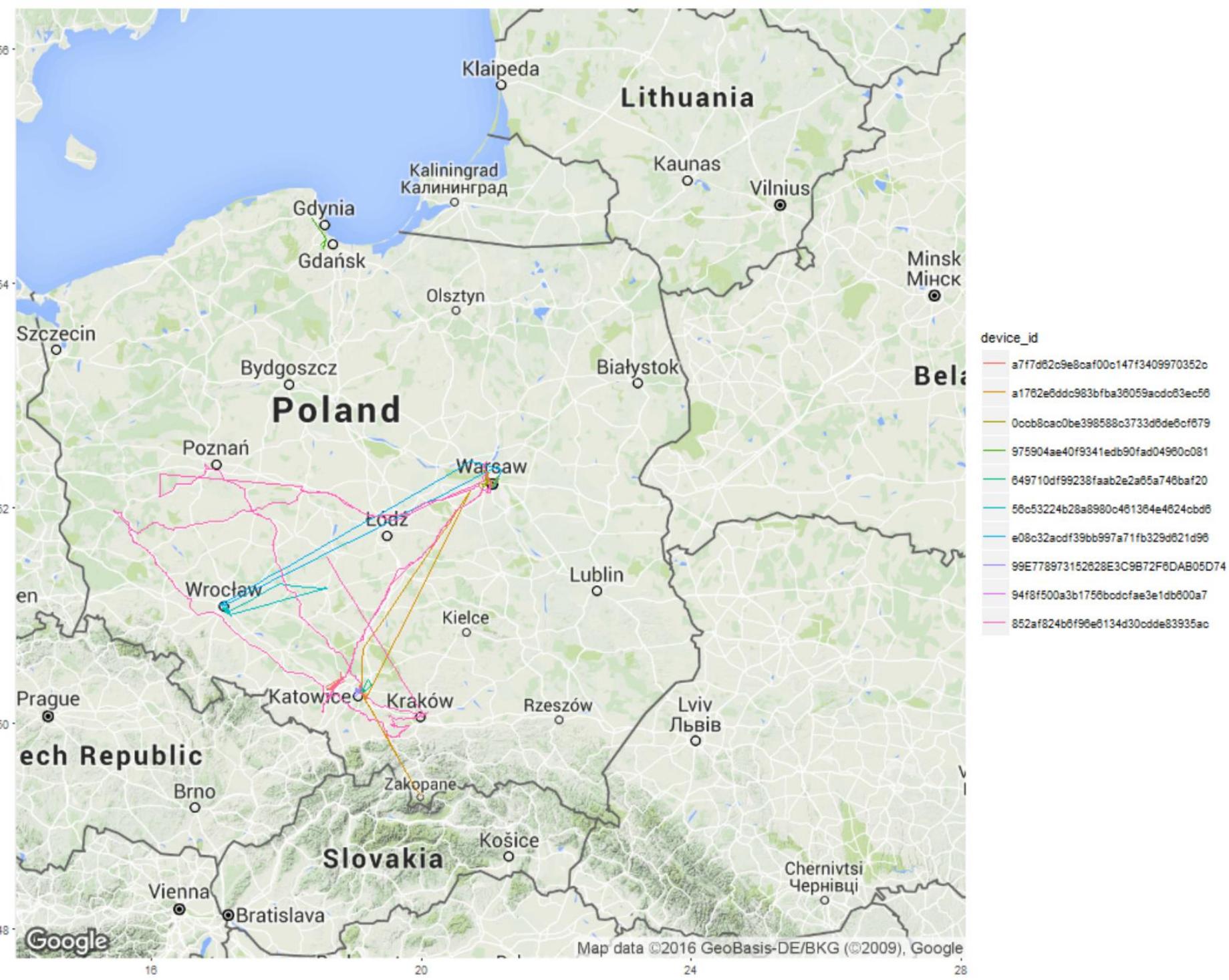


<http://business901.com/blog1/the-misnomer-of-thinking-out-of-the-box/>

Przypadek #4: Swoboda działań

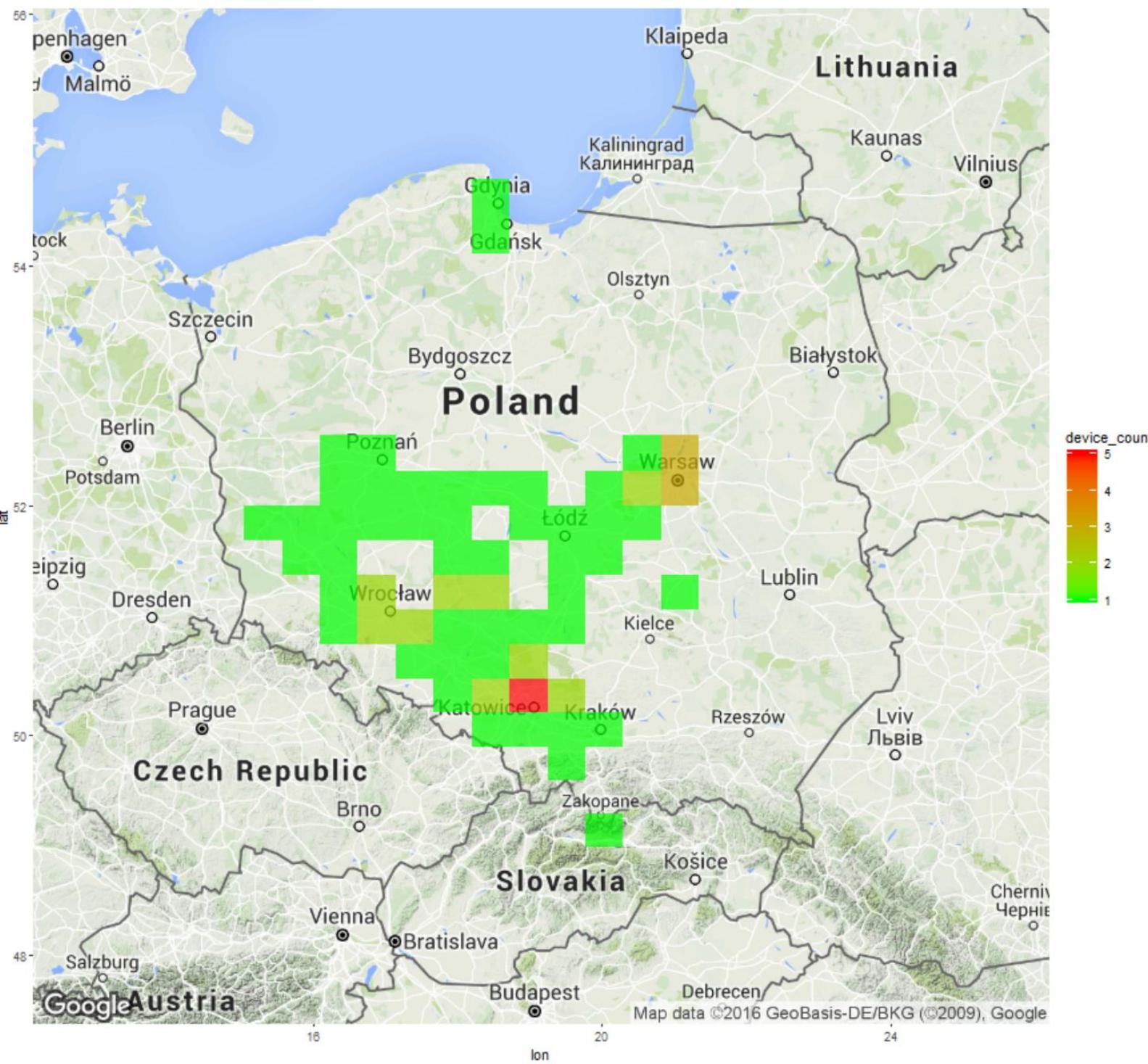
- Duży zbiór danych
- 500 mln wierszy
 - 3 mln sesji
 - 260 tys urządzeń
- Brak mocno sprecyzowanych oczekiwania biznesowych
- Większość przetwarzania zrealizowana w klastrze HDInsight
- Twarda analiza i wizualizacja z użyciem R Servera



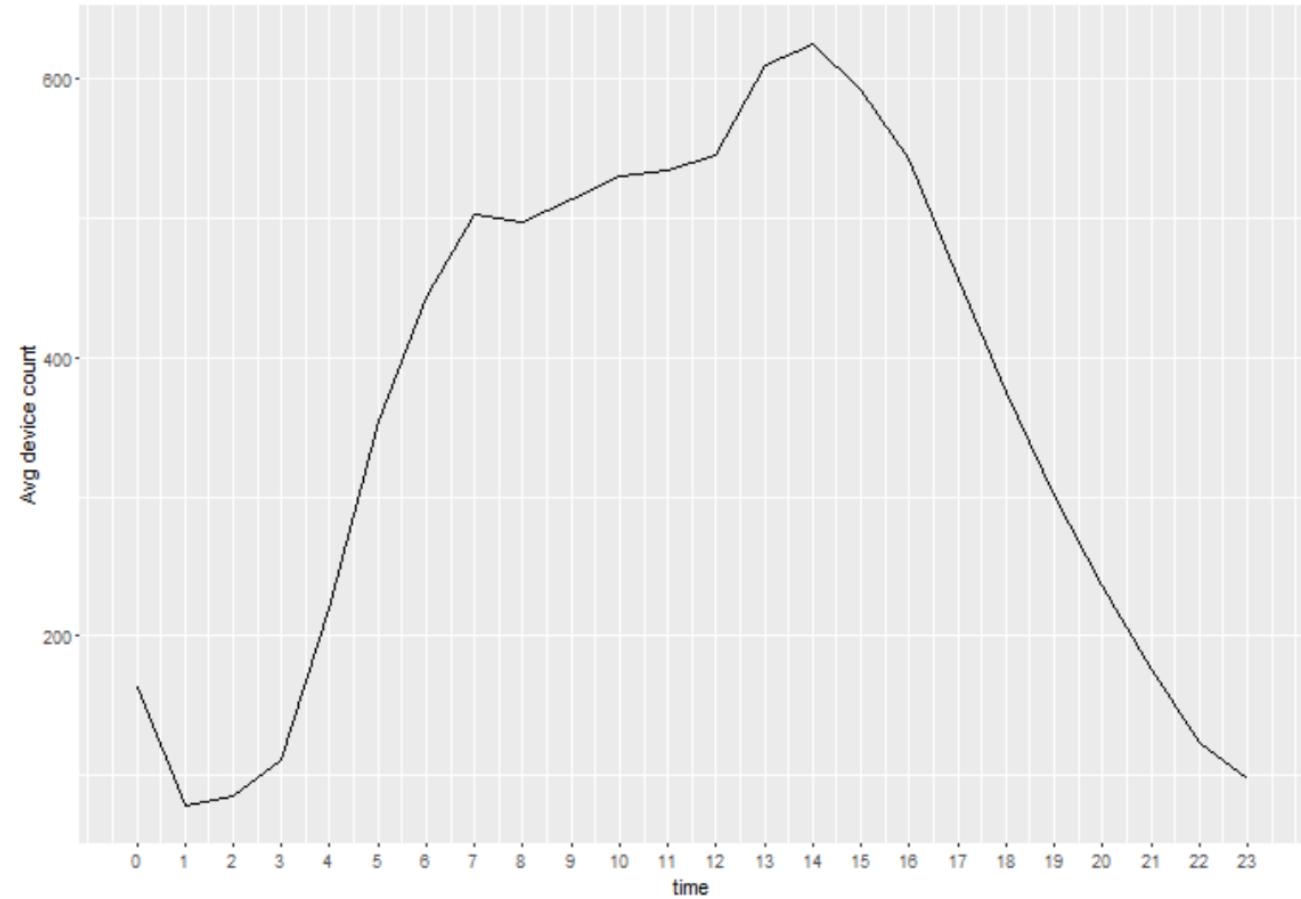
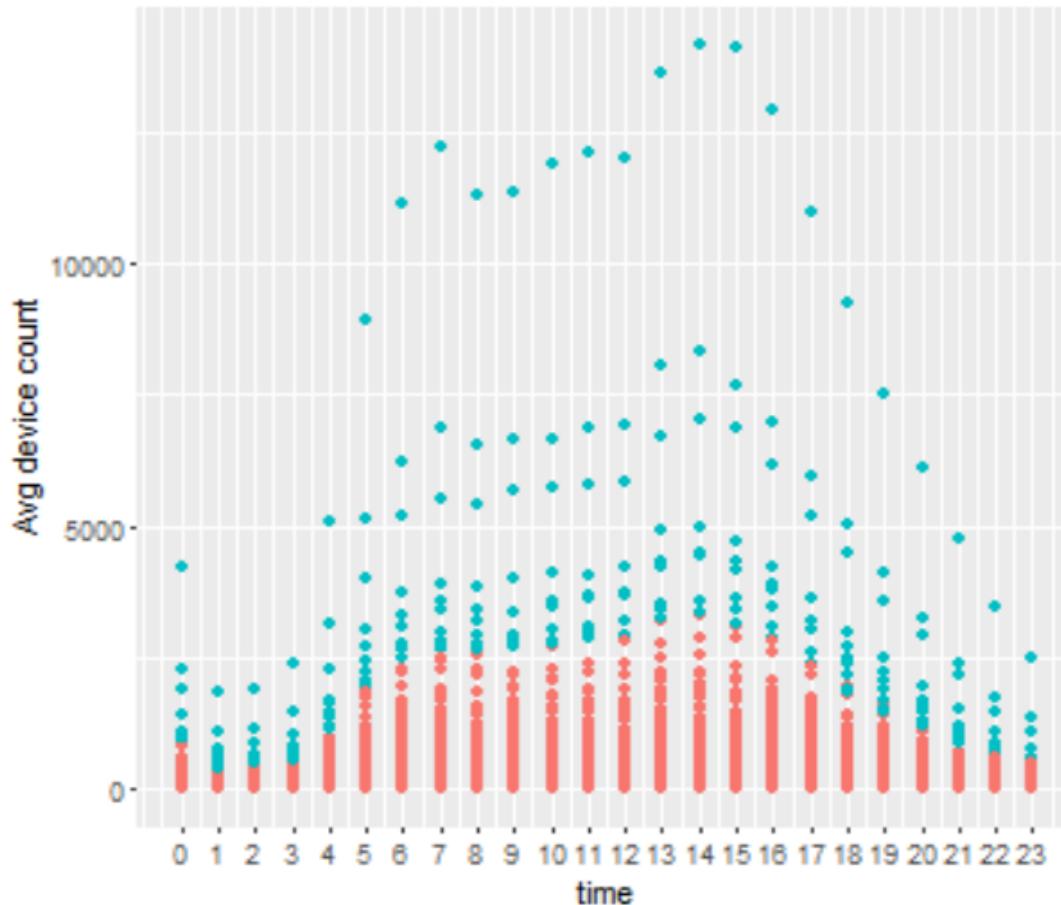


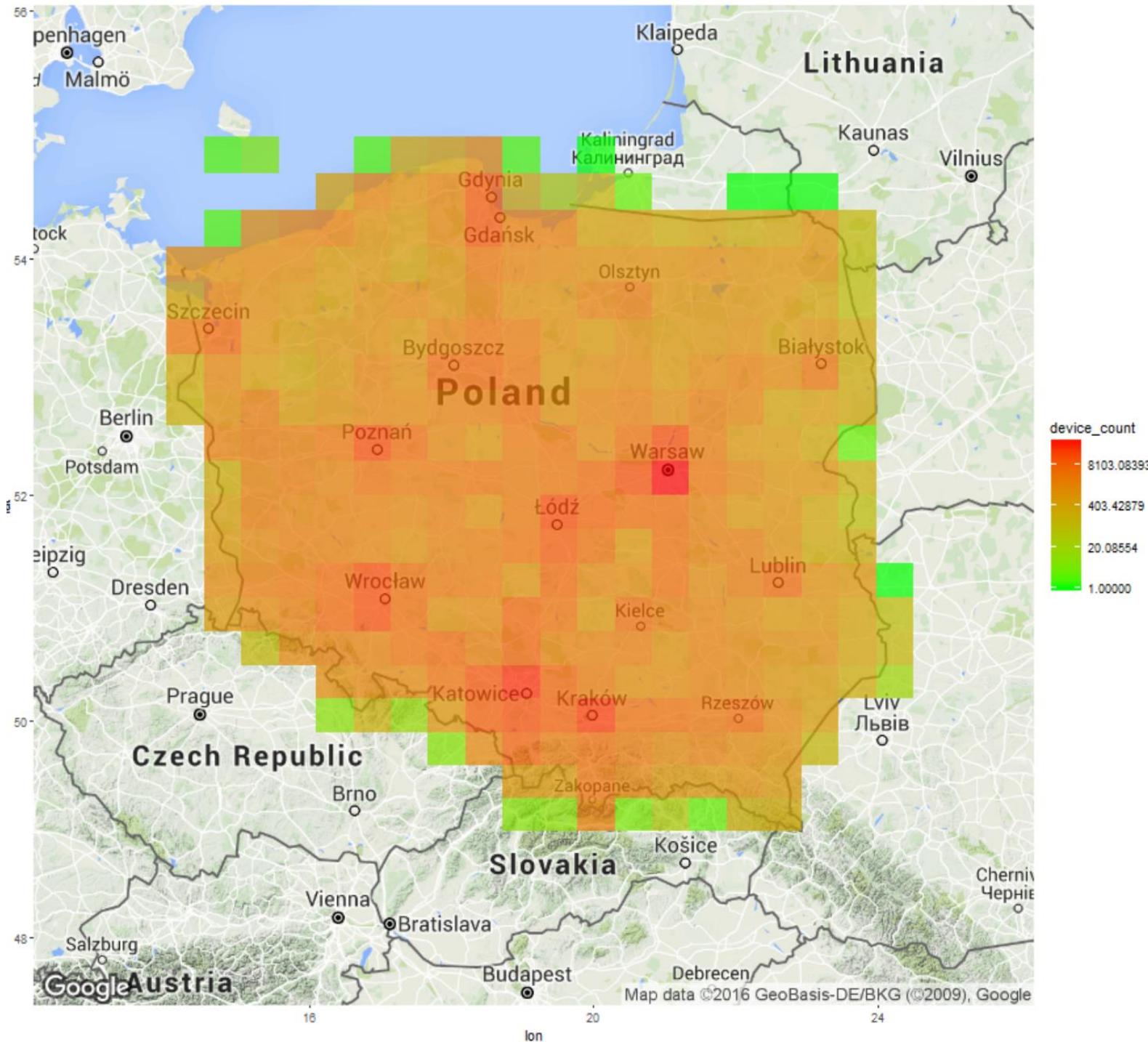






Analiza ruchu godzinowego





Przypadek #4: Rezultaty

- Big data i R znalazły uznanie u klienta
- Szansa na nowe modele biznesowe
- Niektóre z pomysłu gotowe do niemalże natychmiastowego zastosowania (np. dashboardy w Power BI)

Nie zaczynaj
od technologii

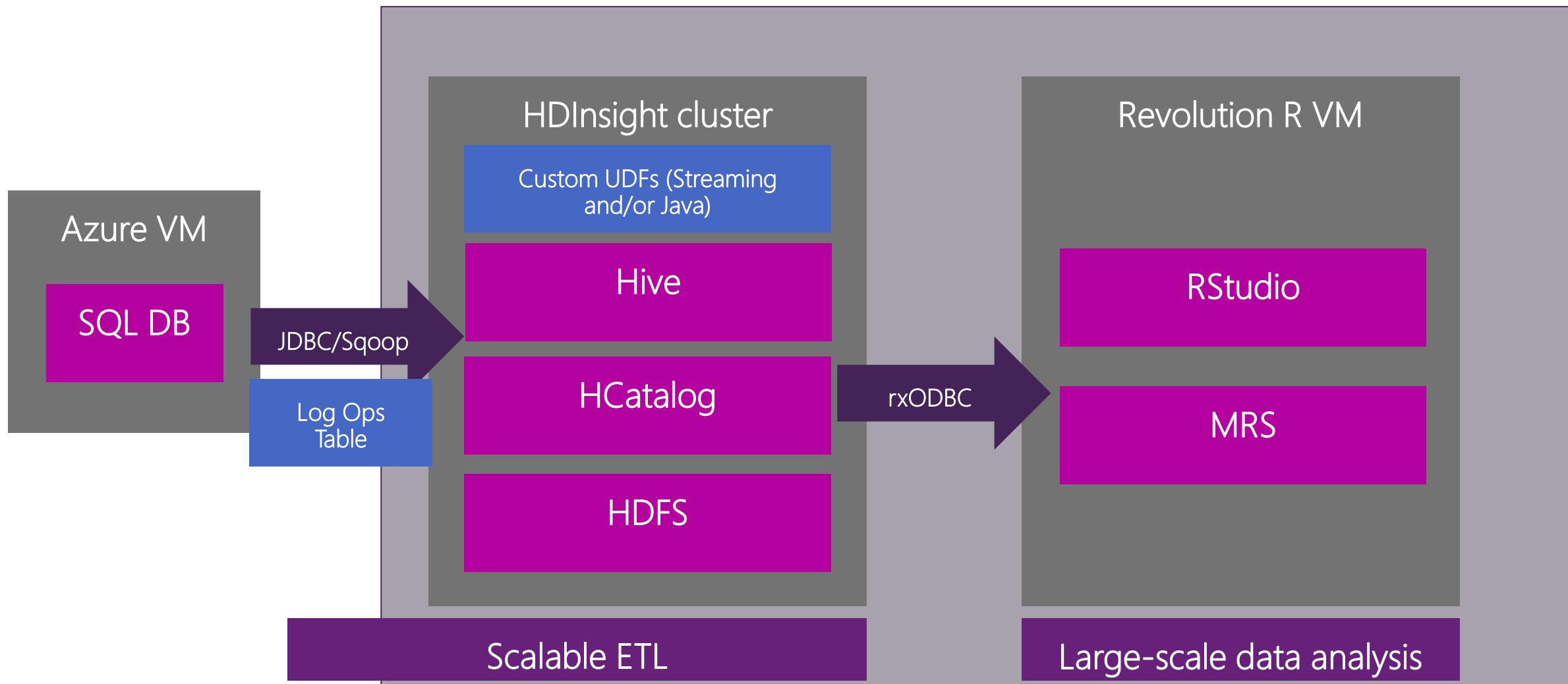


https://commons.wikimedia.org/wiki/File:A_screwdriver_tip.jpg

Przypadek #5: Think big

- Lokalna instancja SQL Servera nie była w stanie poradzić sobie (również wydajnościowo) z nowym typem analiz
- Klient rozważał od pewnego czasu klaster Hadoop jako potencjalną alternatywę
- Mając dostęp do danych źródłowych, mogliśmy pozwolić sobie na umocnienie naszego „business case”

Architektura pilotażowa

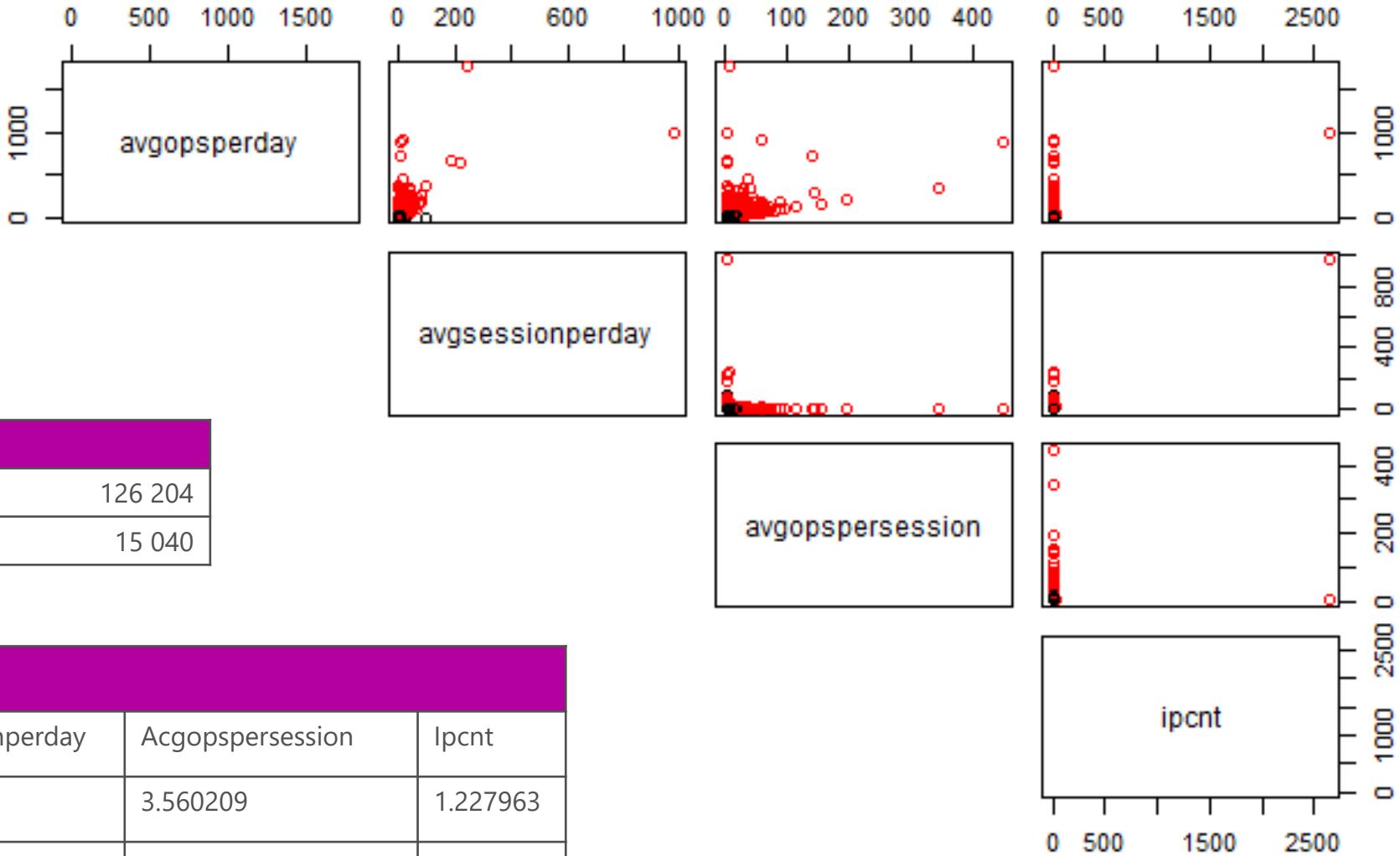


Wykrywanie anomalii za pomocą segmentacji

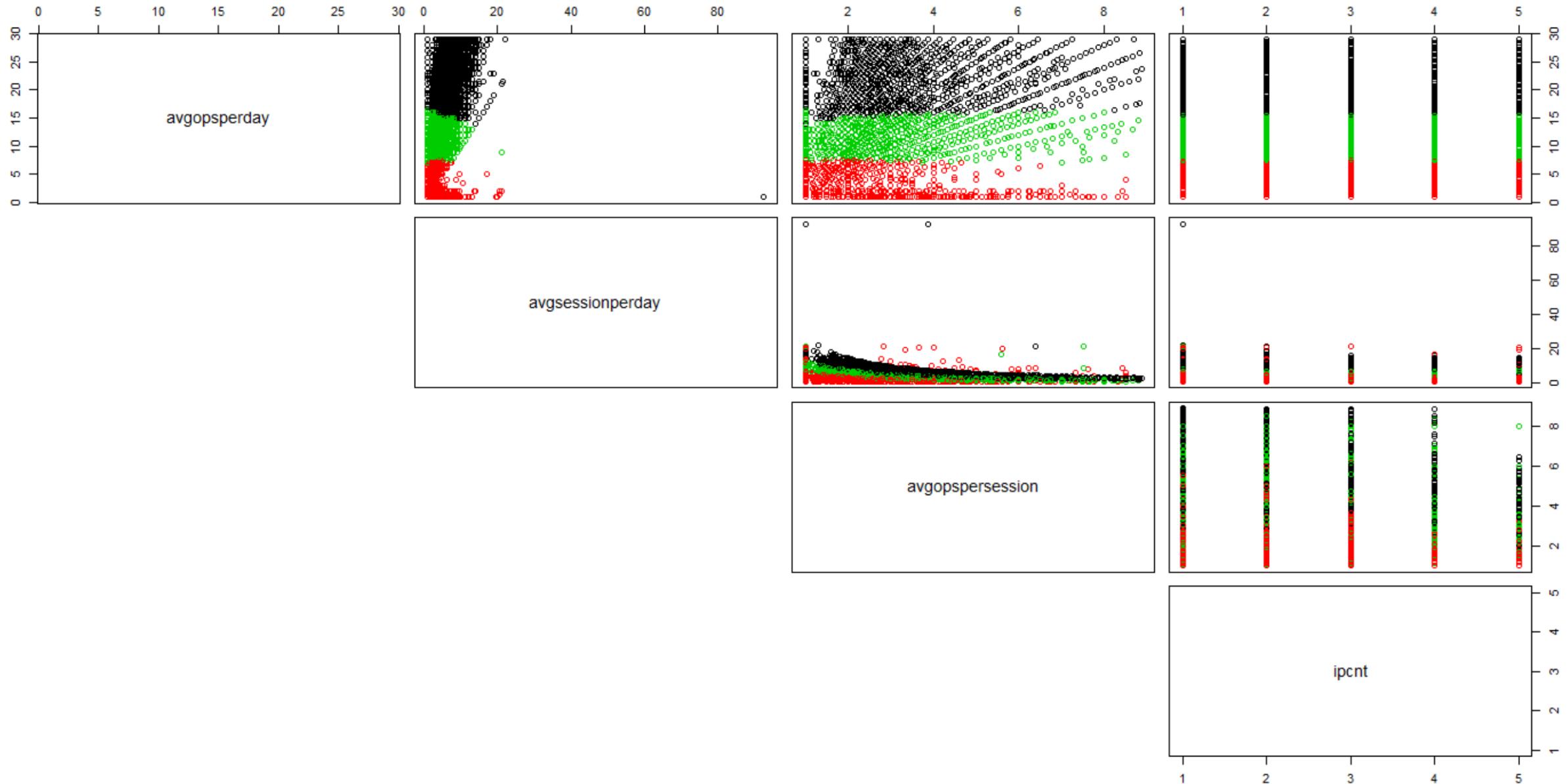
| |
|-----------------------|
| „Skrajni” użytkownicy |
| SYSTEM_IDENTITY1 |
| SYSTEM_IDENTITY2 |
| ANON1 |
| ANON2 |
| ANON3 |

| Rozmiar klastra | |
|-----------------|---------|
| Cluster #1 | 126 204 |
| Cluster #2 | 15 040 |

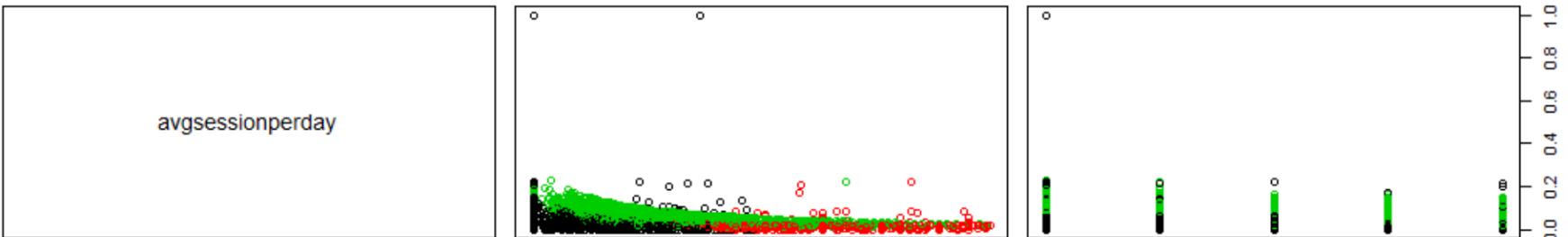
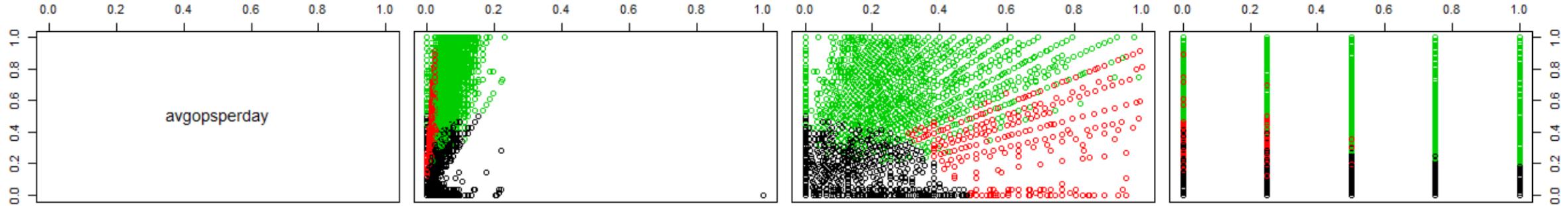
| Położenie klastra | avgopsperday | avgsessionperday | Acgopspersession | Ipcnt |
|-------------------|--------------|------------------|------------------|----------|
| 1 | 7.97499 | 2.380711 | 3.560209 | 1.227963 |
| 2 | 36.81413 | 6.414342 | 7.823085 | 1.640691 |



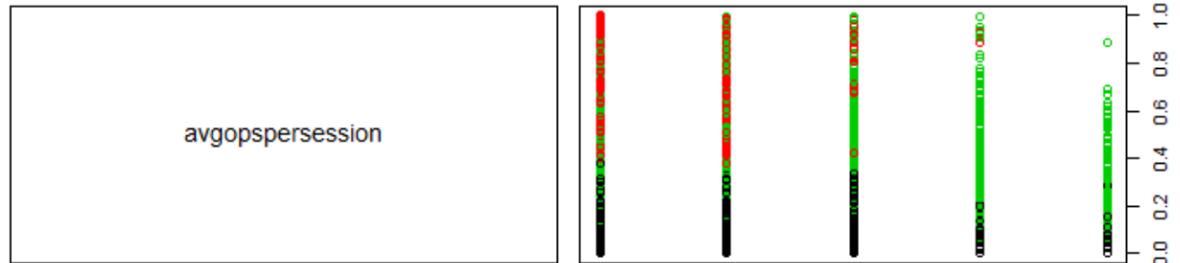
3 segmenty użytkowników [po usunięciu anomalii, bez normalizacji]



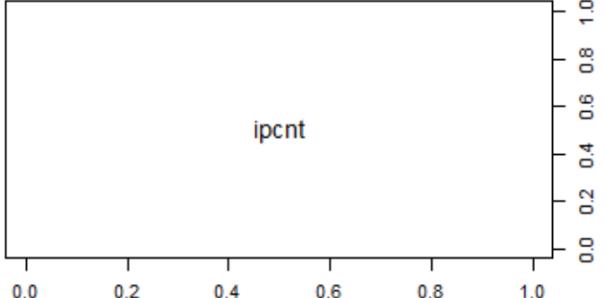
3 segmenty użytkowników [po usunięciu anomalii, z normalizacją]



| Rozmiar klastra | |
|-----------------|-------|
| Cluster #1 | 72820 |
| Cluster #2 | 19956 |
| Cluster #3 | 33487 |

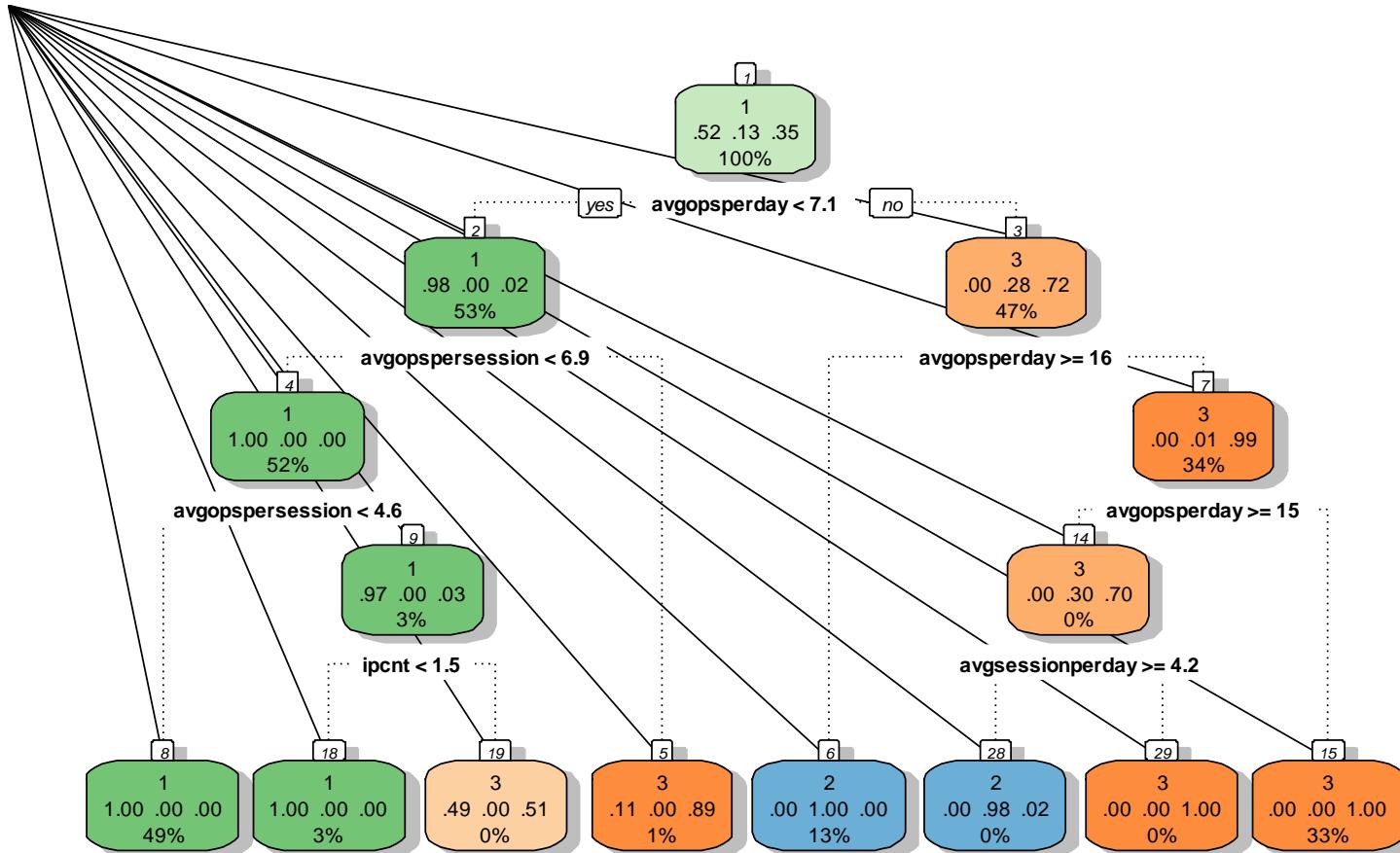


| | Avgopsperday | Avgsessionperday | Avgopspersession | Ipcnt | Segmenty |
|------------|--------------|------------------|------------------|------------|---------------------|
| Cluster #1 | 0.1302338 | 0.01241954 | 0.1521433 | 0.04084386 | Light users |
| Cluster #2 | 0.6132861 | 0.04294849 | 0.3558698 | 0.17688916 | Active+mobile users |
| Cluster #3 | 0.3472406 | 0.01084593 | 0.5794566 | 0.02338967 | Active+static users |



Klasyfikacja użytkowników

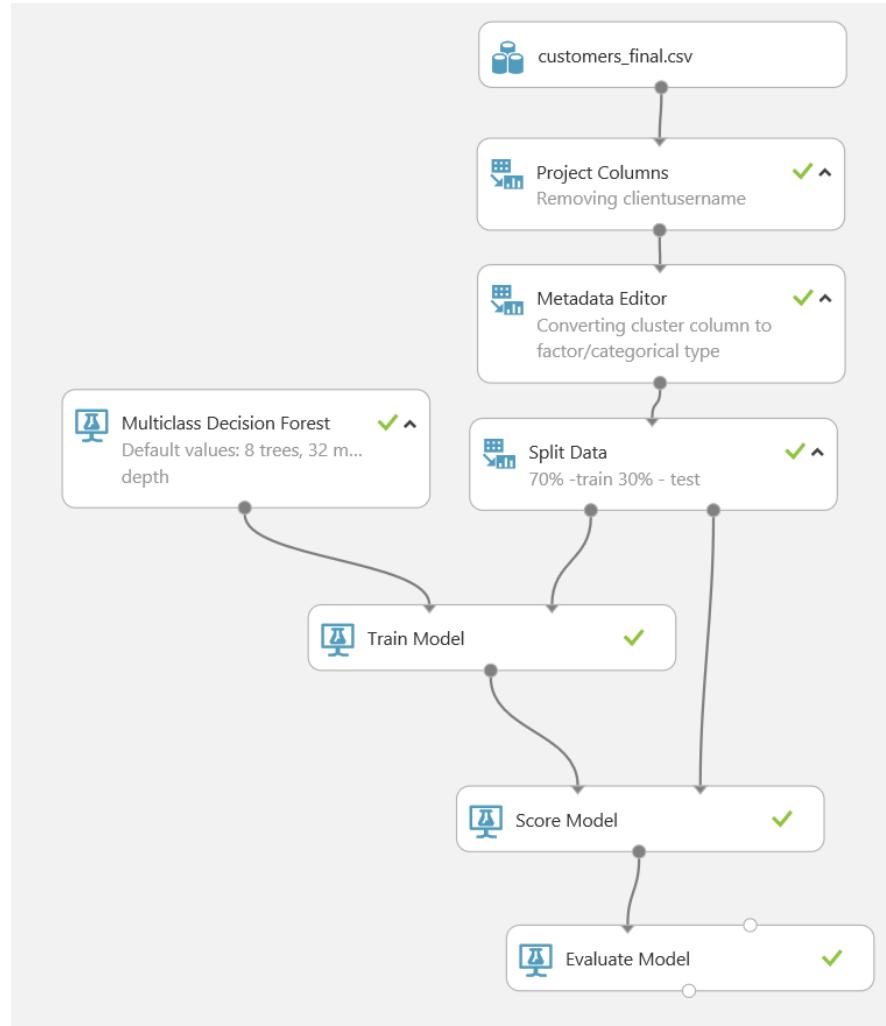
Rezultaty (w oparciu o drzewa decyzyjne):



| | | Confusion Matrix | | |
|------------|----|------------------|----|-----|
| | | Reference | | |
| Prediction | 1 | 2 | 3 | |
| | | 19 684 | 0 | 128 |
| 1 | 3 | 4 933 | 25 | |
| | 24 | 13 080 | | |

Accuracy: 0.9952

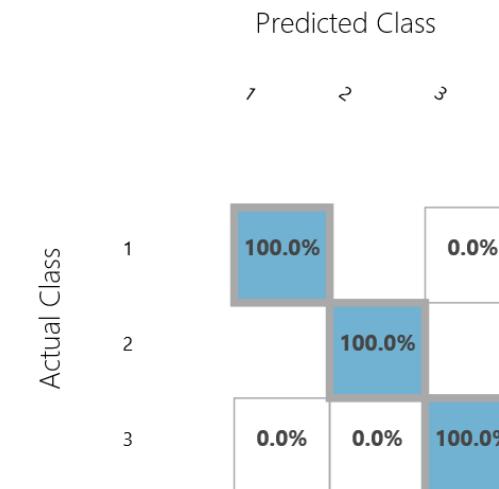
... z wykorzystaniem Azure ML



Metrics

| | |
|--------------------------|----------|
| Overall accuracy | 0.999815 |
| Average accuracy | 0.999877 |
| Micro-averaged precision | 0.999815 |
| Macro-averaged precision | 0.999706 |
| Micro-averaged recall | 0.999815 |
| Macro-averaged recall | 0.999848 |

Confusion Matrix



Przypadek #5: Rezultaty

- Efekt zdecydowanie powyżej oczekiwania klienta
- Dobre spozycjonowanie R Services i R Servera
- Azure niestety nie w produkcji, ale najprawdopodobniej pozostanie środowiskiem test&dev

Nie bój się
porażek

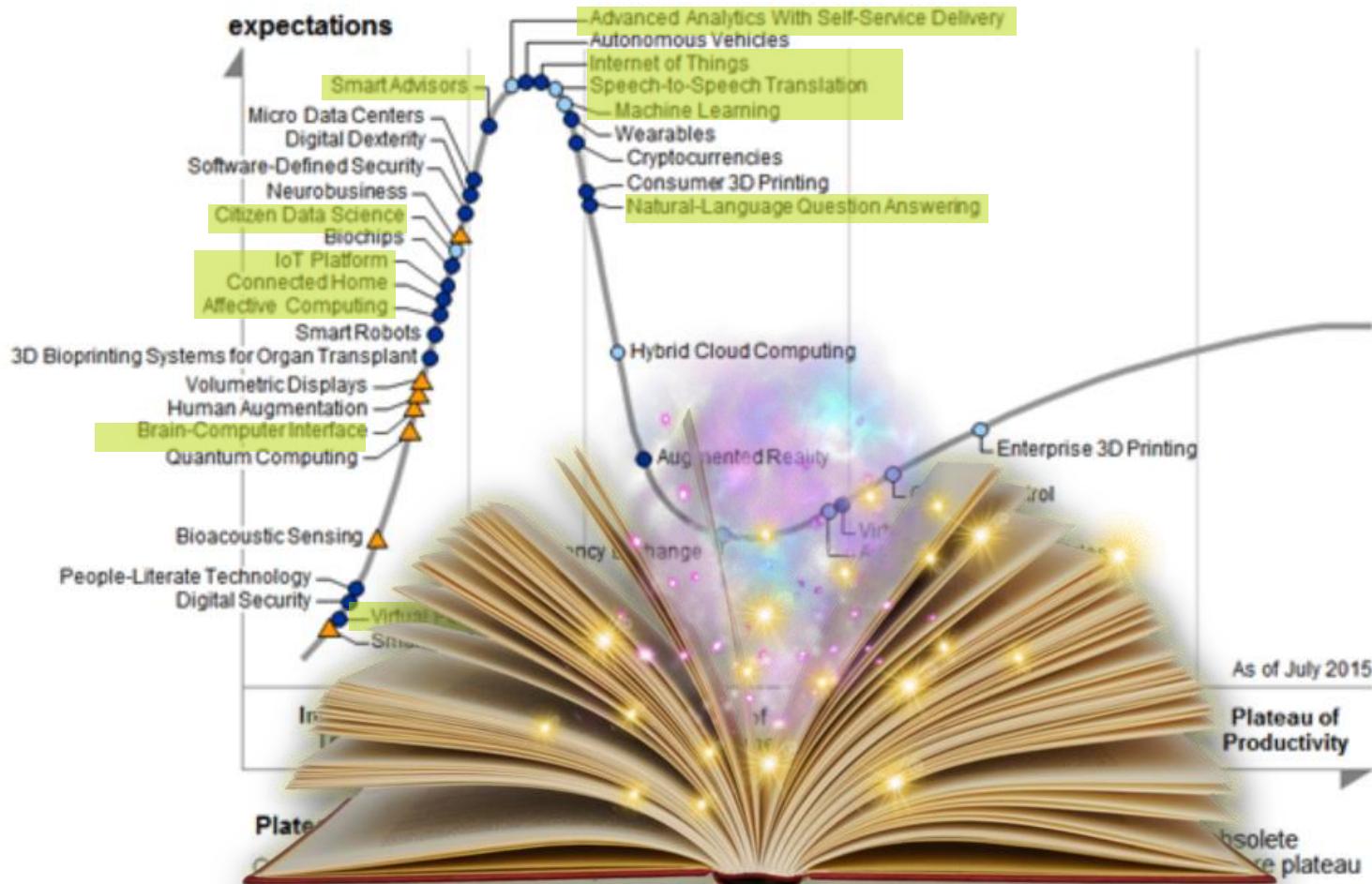


http://www.freeimageslive.co.uk/files/images006/undo_key.jpg

Przypadek #6: Plaza



Najczęstsza przyczyna porażek



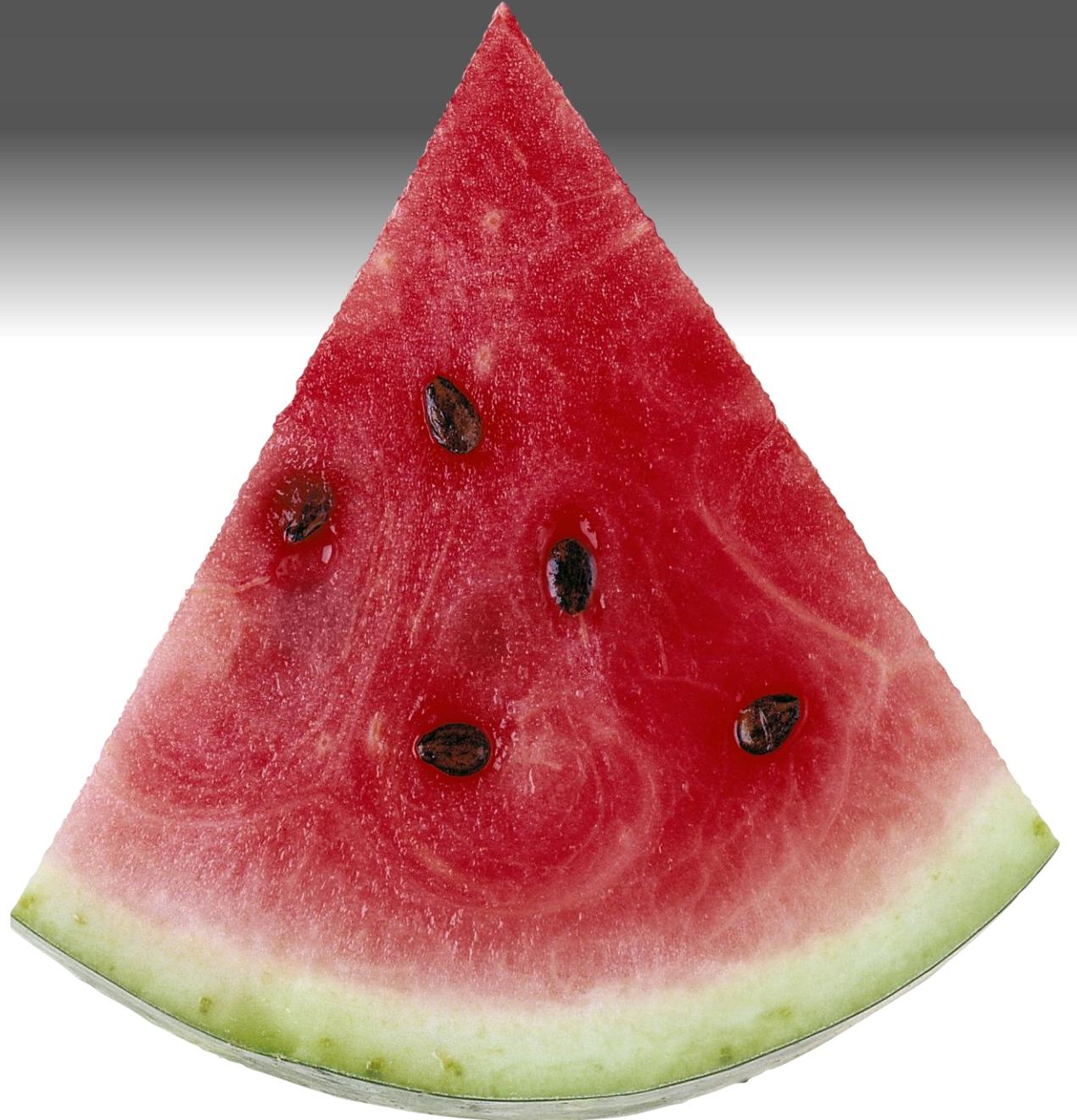
Source: Gartner (August 2015)

Doceniaj ludzi

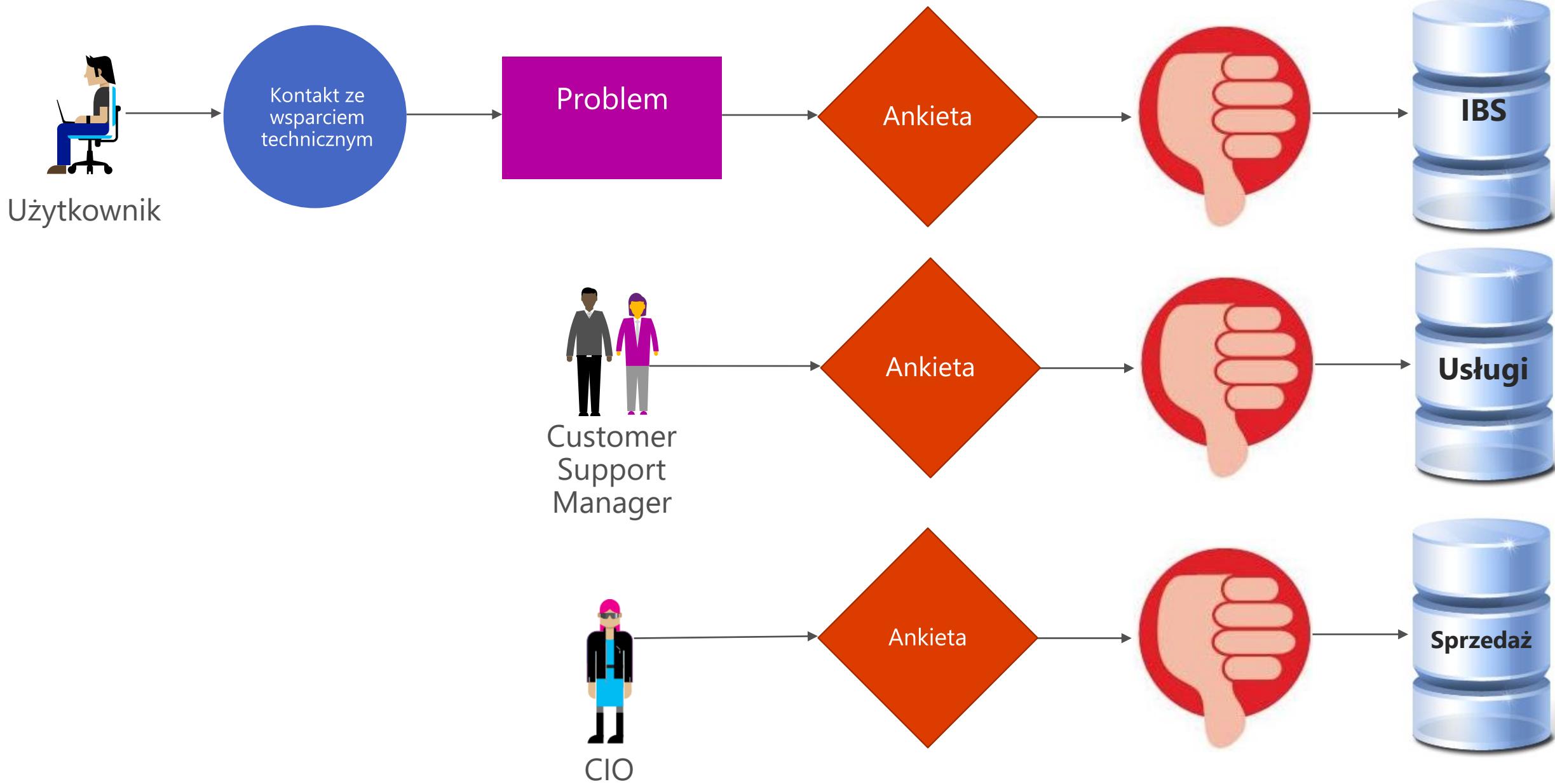


<http://pauldunay.com/your-testing-program-smart-ways-to-get-your-team-on-board/>

Przypadek #7: Arbuz



Dbanie o satysfakcję klienta dziś



Dbanie o satysfakcję klienta już wkrótce

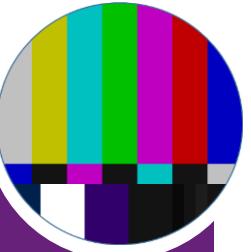
Centralizacja danych



Modele analityczne



Synchronizacja z biznesem



Koniec z silosami

Wspólny zbiór danych pozwoli zunifikować sposób patrzenia na klientów

Uczenie maszynowe

Krótszy czas potrzebny na zmianę

Rozmowa

Nawiązanie dialogu pomiędzy różnymi interesariuszami

Satysfakcja klienta w przyszłości



Bogatszy arsenał
narzędzi do
zbierania danych



Opinie zbierane i
analizowane na
bieżąco



Jednolity
wizerunek firmy

Machine Learning University

ML Learning Paths: Curated Collections of ML-Related Content

- Ścieżki szkoleniowe przygotowane przez ekspertów dziedzinowych
 - [Introduction to Machine Learning Concepts and Uses](#). A series of resources for people with no background in machine learning
 - [Machine Learning Fundamentals, Part I](#). A strong foundation on which you can successfully build your ML knowledge. Activities include an illustrative talk, an essential textbook, and the first 10 lectures of a CalTech ML course.
 - [Machine Learning Fundamentals, Part II](#). Advanced lectures from the CalTech ML course that explore a variety of topics, including overfitting, regularization, support vector machines, and more
- Zróżnicowany poziom trudności
 - [Internal Machine Learning Course - Beginner](#). An internal course covering an introduction to ML, linear representations and learning, and linear learning tools such as TLC (the Microsoft machine learning toolkit) and Vowpal Wabbit
 - [Internal Machine Learning Course - Intermediate](#). An internal course covering decision trees, ensemble models, metrics and advanced learning problem types, and a TLC tutorial
 - [Internal Machine Learning Course - Fast ML and Clustering](#). An internal course covering online learning, feature hashing, parallel learning, OWL-QN on Cosmos, TLC on Cosmos, and clustering
 - [Internal Machine Learning Course - Interaction](#). An internal course covering offline evaluation, online exploration, and active learning
- Bazuje na zasobach wewnętrznych i zewnętrznych
 - [Internal Machine Learning Course - Modeling](#). An internal course covering model-based machine learning and deep learning
 - [Bayes Confusion Matrix](#). An introduction to the fundamentals of data science through the lens of Bayesian probability theory

Ciągłe doszkalanie

- Szkolenia dostępne dla wszystkich pracowników
- Zróżnicowany poziom zagadnień
- Minimalny próg wejścia
- Dostęp do archiwaliów



About

Submit an Idea

- Cykliczne, praktyczne wyzwania dla chętnych
- Scenariusze mocno osadzone w rzeczywistości
- Interdyscyplinarne zespoły

Next hackathon: January 12th 2015
Details coming soon

Recent
Hackathons

O365 Support
Hackathon
September 2015

Display Ads
Hackathon
June 2015

Forecast
Hardware
Purchases
at the
Microsoft
Store Online
March 2015

Xbox Holiday
Offer Hackathon
October 2014

learn more →

learn more →

learn more →

learn more →

7 zasad zwiększających prawdopodobieństwo sukcesu

- Kieruj się prostotą
- Ufaj chmurze
- Sprzedawaj historie
- Inspiruj
- Nie zaczynaj od technologii
- Nie bój się porażek
- Doceniaj ludzi

Co dalej?

Co oferujemy:

- Advanced Analytics Labs (2 dni)
 - Machine Learning, Stream Analytics, Spark/HDInsight, Data Factory, Power BI
- R Labs (1-2 dni)
 - R Server, R Services
- (Bezpłatne) zaangażowanie w indywidualne projekty

Pytania?

