



# Inteligencja biznesowa (BI)

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie  
AGH University of Science and Technology

2023/2024

<https://teams.microsoft.com/l/team/19%3a-024Z7S12Bw-RmESiajWuRg5eOdbk0VKTDGfEaiU2ns1%40thread.tacv2/conversations?groupId=d3e61234-4dea-42ed-b643-25c3b11541ad&tenantId=80b1033f-21e0-4a82-bbc0-f05fdcccd3bc8>

## Wprowadzenie

Dlaczego organizacje inwestują w BI

Definiowanie komponentów technologicznych używanych w rozwiązaniach BI

Odkrywanie aktywności w codziennym BI

Wprowadzenie BI do własnej kariery

Poznasz wartość inteligencji biznesowej i umiejętności wymagane, aby być wystarczającym.

## Wprowadzenie

Dlaczego organizacje inwestują w BI

Definiowanie komponentów technologicznych używanych w rozwiązaniach BI

Odkrywanie aktywności w codziennym BI

Wprowadzenie BI do własnej kariery

Poznasz wartość inteligencji biznesowej i umiejętności wymagane, aby być wystarczającym.

Work hard vs  
Work smart vs  
Work enough



## Rozumienie tego, czym jest inteligencja biznesowa

Dlaczego organizacje inwestują w inteligencję biznesową (BI)

Obszary wchodzące w skład BI

Działania zachodzące w pracy związanej z BI

## BI jako ścieżka swojej ścieżki kariery

Struktury organizacyjne BI

Działania zarządcze mające na celu rozpoczęcie budowy i włączania umiejętności BI do swoich zespołów

Działania, które należy podjąć, aby rozwijać swoją ścieżkę kariery i umiejętności związane z BI

## BI Zapewnia wgląd w dane



Mówimy o wartości analityki biznesowej, ponieważ technologia rozwinęła się w taki sposób, że firmy mogą podejmować bardziej świadome decyzje, a nie kierować się przypuszczeniami.

Dzięki analityce biznesowej firmy mogą teraz odpowiadać na pytania strategiczne i operacyjne za pomocą danych. Mogą odkryć, dlaczego sprzedaż rośnie lub spada tylko w określonym obszarze ich rynku. Firmy mogą przewidzieć, czy będą miały nadmiar lub niedobór zapasów na wcześniejszym etapie swoich procesów, aby mieć czas na dostosowanie i skorygowanie tego.

Source: Jamie Champagne

Wszystko to wynika z rozwoju postępu technologicznego, choć niestety często sprawia wrażenie przeciążenia danymi.

## Data Overload

W erze cyfrowej, dane są produkowane, gromadzone i przetwarzane w niespotykanej dotąd skali, prowadząc do sytuacji, w której jednostki i organizacje mogą czuć się przytłoczone przez ilość dostępnych informacji.



Może to także wpływać negatywnie na produktywność, zdolność do skupienia się i ogólne samopoczucie. Przeciążenie danymi może prowadzić do paraliżu decyzyjnego, gdyż trudno jest wyodrębnić istotne informacje z morza dostępnych danych.

<https://martech.org/wp-content/uploads/2014/08/data-information-overload-ss-1920.jpg>

## Data Overload (przyczyny, skutki)

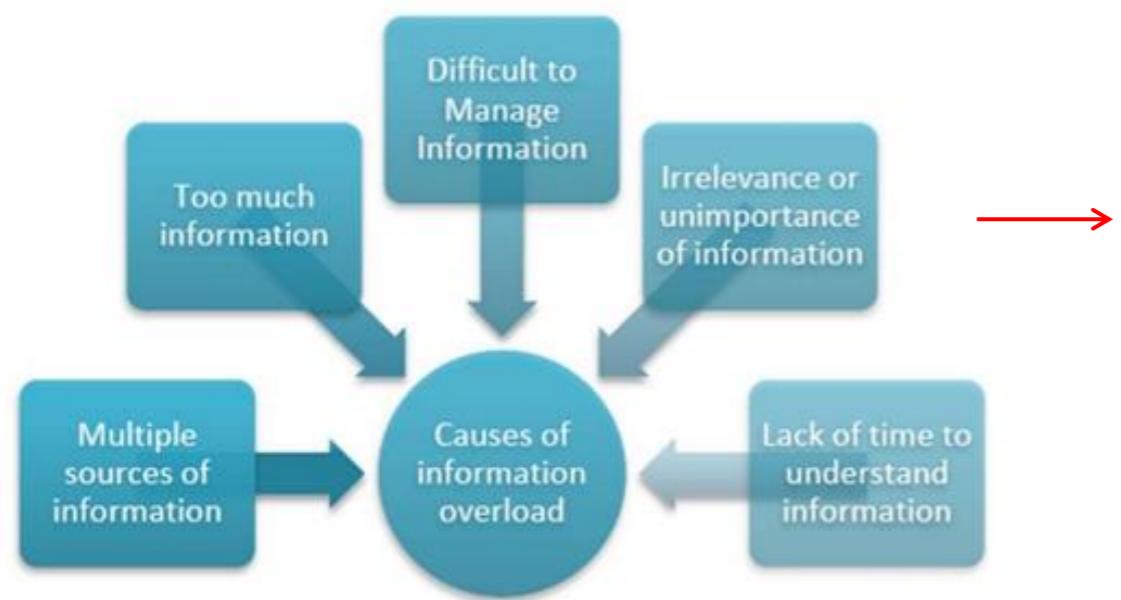


Fig. 2 Factors cause information overload

- Słaba koncentracja z powodu przeładowania pamięci krótkotrwałej
- Choroba pośpiechu, która jest przekonaniem, że należy ciągle ścigać się z czasem
- Wszechobecna wrogość prowadząca do przewlekłego stanu drażliwości bliskiej gniewowi lub nawet furii
- Przyzwyczajenie lub nadmierna stymulacja, która powoduje, że mózg się wyłącza i wchodzi w stan podobny do transu
- Kompulsja "bycia podłączonym" jest silną potrzebą sprawdzania e-maili, poczty głosowej i Internetu w celu pozostania "w kontakcie,"
- Tradycyjny stres, w tym obniżona odporność, zaburzenia równowagi endokrynologicznej, depresja i doświadczenie "wypalenia"

*Effect of Information Overload on Decision's Quality, Efficiency and Time: International Journal of Latest Engineering Research and Applications (IJLERA) ISSN: 2455-7137*

## Data Overload – ćwiczenie (notatki + dyskusja) – 15 min

Podaj przykład Data Overload w twoim najbliższym otoczeniu

Co na niego się składa – skąd pochodzi

Co na niego wpływa – jakie są przyczyny

Jakie są objawy

Jakie są skutki

Jakie narzędzia wypracowałeś/wypracowałaś aby sobie z tym poradzić ?

Np. Płatności, Rolki (TT/SM), Maile vs Kalendarz

## Data Overload – ćwiczenie (notatki + dyskusja) – 15 min

Podaj przykład Data Overload wyobrażasz sobie, że zachodzi w środowisku organizacji

Co na niego się składa – skąd pochodzi

Co na niego wpływa – jakie są przyczyny

Jakie są objawy

Jakie są skutki

Jakie narzędzia można wypracować aby sobie z tym poradzić ?

Np. Zmiany w procesie change managementu, HR Pipeline, Estymacja wartości produktów

## Data Overload

W odpowiedzi na te wyzwania, coraz więcej uwagi poświęca się narzędziom i technikom zarządzania danymi, takim jak data mining, analiza big data i sztuczna inteligencja, które mają na celu pomóc w filtrowaniu, analizie i wykorzystywaniu danych w bardziej efektywny sposób



## Strategie biznesowe i technologie do analizy danych w celu podejmowania świadomych, decyzji biznesowych opartych na danych.



Następnie organizacja zaczyna eksplorować te dane, aby uzyskać lepszy wgląd i udostępnić więcej tych danych, a następnie dzięki tym spostrzeżeniom zaczynamy tworzyć wizualizacje danych.

Przyspiesza to procesy decyzyjne, co z kolei prowadzi do opracowania strategii zarządzania danymi

Graphic: Jamie Champagne



## Business Intelligence (BI)

- Data analysis**
- Data management and warehousing**
- Data transformation**
- Big Data**
- Reporting and dashboards**
- Online Analytical Processing (OLAP)**
- Data and process mining**
- Benchmarking**
- Predictive and prescriptive analytics**

Obejmuje to takie obszary jak:

- analiza danych
- zarządzanie danymi i hurtownie danych
- transformacja danych
- big data - dane o większej różnorodności, objętości i szybkości, co z kolei oznacza przeciążenie danymi
- raportowanie i pulpity nawigacyjne, a w tym celu często korzysta się z funkcji przetwarzania analitycznego online (OLAP) oraz ogólnej eksploracji danych i procesów, aby zrozumieć, jak działają analizowane fakty
- analiza porównawcza (benchmarking) jest kluczem do wiedzy branżowej
- BI wykorzystywany też jest do przeprowadzania analiz predykcyjnych, a także analiz preskryptywnych, które dotyczą naszych przyszłych działań (*analizy te pomagają firmom patrzeć na przyszłość i analizować je z odpowiednią dokładnością. Zdolność ta zawsze była ważna – ale nigdy nie była tak krytyczna jak teraz*) - analizy są często określane jako „ostatnia faza [analiz biznesowych](#)”. Jest to również najbardziej złożony i stosunkowo nowy – obecnie znajdujący się na szczycie [cyklu Hype dla analiz i analiz biznesowych firmy Gartner 2020](#).

Graphic: Jamie Champagne

Obejmuje to takie obszary jak:

- analiza danych
- zarządzanie danymi i hurtownie danych
- transformacja danych
- big data - dane o większej różnorodności, objętości i szybkości, co z kolei oznacza przeciążenie danymi
- raportowanie i pulpity nawigacyjne, a w tym celu często korzysta się z funkcji przetwarzania analitycznego online (OLAP) oraz ogólnej eksploracji danych i procesów, aby zrozumieć, jak działają analizowane fakty
- analiza porównawcza (benchmarking) jest kluczem do wiedzy branżowej
- BI wykorzystywany też jest do przeprowadzania analiz predykcyjnych, a także analiz preskryptywnych, które dotyczą naszych przyszłych działań (*analizy te pomagają firmom patrzeć na przyszłość i analizować je z odpowiednią dokładnością. Zdolność ta zawsze była ważna – ale nigdy nie była tak krytyczna jak teraz*) - analizy są często określane jako „ostatnia faza [analiz biznesowych](#)”. Jest to również najbardziej złożony i stosunkowo nowy – obecnie znajdujący się na szczycie [cyklu Hype dla analiz i analiz biznesowych firmy Gartner 2020](#).

- At the Peak
  - Decision Intelligence
  - Continuous Intelligence
  - Natural Language Generation (NLG)
  - Edge Analytics
  - Data Storytelling
  - Digital Ethics
  - Explainable AI
  - Data Catalog
  - Prescriptive Analytics



Sliding Into the Trough

[Technology](#)   [Insights](#)   [Expert Guidance](#)   [Tools](#)   [Connect with Peers](#)

**Hype Cycle for Analytics and Business Intelligence, 2020**

Graphic: Jamie Champagne

# Hype Cycle for Artificial Intelligence, 2023



gartner.com

Source: Gartner  
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. 2079794

Gartner®

<https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle>

# Hype Cycle for Artificial Intelligence, 2023



Dzielony jest na pięć stadiów:

**innovation trigger** – “spust innowacyjności”, czyli początek istnienia nowej technologii i zainteresowania wokół niej; wszyscy o tym mówią – mało kto widział to na własne oczy

**peak of inflated expectations** – “szczyt zawyżonych oczekiwania”, czyli moment gdy wyobrażenia i oczekiwania na temat danej technologii uzyskują najwyższy poziom; jeśli wokół technologii narasta jakąś bańka spekulacyjna, to to jest ta chwila gdy za chwilę pęknie...

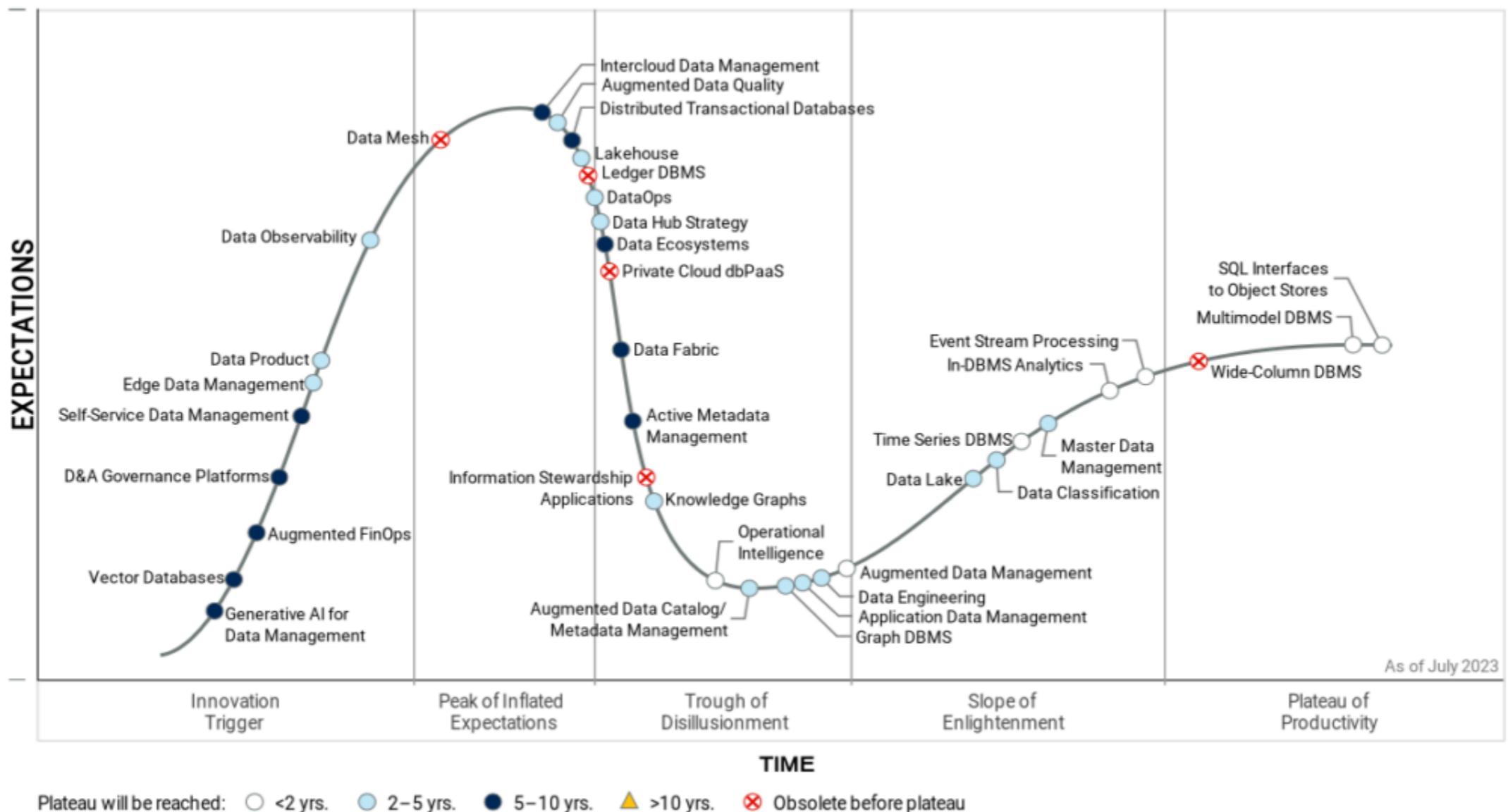
**trough of disillusionment** – “koryto rozczarowania”, czyli zderzenie oczekiwania z ograniczeniami i wadami danej technologii; nagle okazuje się, że technologia jest beznadziejna, nie warto się nią zajmować, a wiele firm, które zaczęły się w niej specjalizować – upada

**slope of enlightenment** – “krzywa oświecenia”, czyli stopniowy powrót użytkowników do danej technologii, pogodzonych z jej ograniczeniami i wadami; abstrahując od problemów, okazuje się jednak, że technologia jest dość przydatna w określonych obszarach i warto w nią inwestować czas i pieniądze

**plateau of productivity** – “płaskowyż produktywności”, czyli wykorzystywanie danej technologii w codziennych procesach wytwórczych.

Na krzywej naniesione są jej szczegółowe dziedziny, z oznaczeniem w jakim okresie dane stadium zostanie osiągnięte: mniej niż 2 lata, 2 do 5 lat, 5 do 10 lat oraz powyżej 10 lat.

Hype Cycle for Data Management, 2023



Gartner

## Hype Cycle for Big Data, 2013

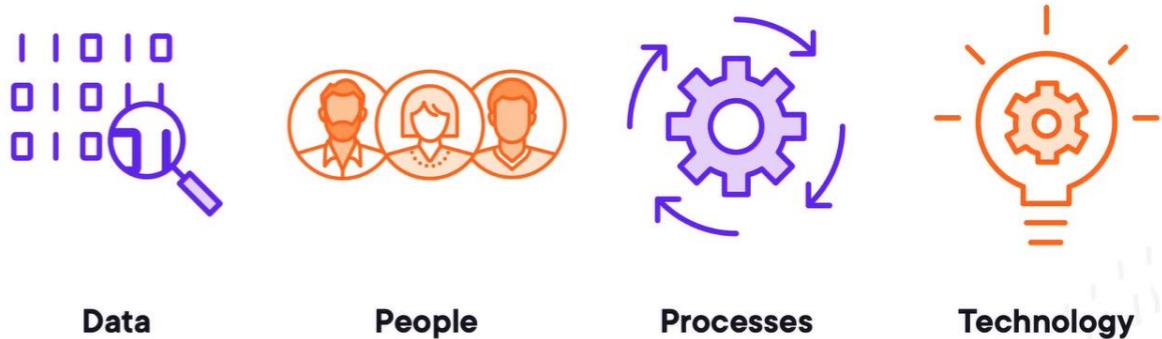


Source: Gartner

- Myśląc o analityce biznesowej, często można podzielić ją na cztery elementy, które budują strategię lub portfolio analityki biznesowej.

- Oczywiście są to **dane** (dane napędzają)
- Następnie należy wziąć pod uwagę **ludzi**, którzy analizują dane, którzy je czyszczą, którzy je organizują, którzy interpretują dane, i są to te same osoby, które często mogą pomóc w zdefiniowaniu **zasad i procesów**, które pozwalają nam wykorzystać te spostrzeżenia.
- Do tego dochodzi **technologia automatyzująca** pracę z danymi, a także infrastruktura wspierająca potrzeby użytkowników biznesowych.

### Business Intelligence Elements



Graphic: Jamie Champagne

Dlatego właśnie organizacje inwestują w całe zespoły BI i opracowują strategie BI, w zakresie analityki biznesowej.

Strategia Business Intelligence ma na celu połączenie perspektywy biznesowej z danymi poprzez wykorzystanie technologii. Zwykle inteligencja biznesowa w organizacji jest napędzana celami biznesowymi, takimi jak zdefiniowanie wskaźników, które firma chce lub musi osiągnąć, oraz ustanowienie standardów, kluczowych wskaźników wydajności, tych kluczowych wskaźników wydajności i licznych celów biznesowych do osiągnięcia. Sposobem na osiągnięcie tego celu jest uzyskanie wglądu w dane.



**Business Goals**  
**Defining business metrics and key performance indicators (KPIs)**



**Data Insights**  
**Data tools, technologies, and structure and governance in place to leverage the data**

Graphic: Jamie Champagne

## **BI odpowiada na potrzeby biznesowe, a budowanie programu BI obejmuje**

- pozyskiwanie danych i zarządzanie nimi
- elementy, które są zawarte w ETL, czyli ekstrakcję tych danych, transformację danych i ładowanie danych, aby można je kolejno było wizualizować
- postrzeganie danych z nowych perspektyw zwiększa istotność obszaru jakości danych i zarządzania dostępami
- role, kompetencje potrzebne po stronie danych wraz ze szkoleniem i zrozumieniem po stronie biznesowej, aby wykorzystać pracę z danymi w celach inwestycyjnych



### **BI Program**

**Data sourcing and management**

**ETL (Extract, transform, and load)**

**Visualizations and dashboards**

**Data quality and governance**

**Data security and privacy**

**Data and business roles**

*Graphic: Jamie Champagne*

## BI odpowiada na potrzeby biznesowe



Programy\* BI mogą nie tylko odpowiadać na pytania biznesowe, ale także napędzać działania biznesowe.

*\*zorganizowany zestaw powiązanych projektów i działań, które są zarządzane w sposób koordynowany, aby osiągnąć cele strategiczne i korzyści, których nie można by osiągnąć, gdyby projekty były zarządzane oddzielnie. Program skupia się na osiąganiu celów strategicznych organizacji poprzez efektywne wykorzystanie zasobów, harmonizację celów poszczególnych projektów z ogólną strategią firmy, oraz zapewnienie spójności i synergii między projektami.*

Graphic: Jamie Champagne

## BI odpowiada na potrzeby biznesowe

Czy powinniśmy inwestować w wewnętrzne zasoby czy outsourcing?



*Analiza wskaźników efektywności pracowników i zarządzanie talentami, aby optymalizować wydajność i motywację zespołu*

*Identyfikacja wzorców i trendów w danych dotyczących zatrudnienia, które mogą wskazywać na potrzebę zmian w polityce personalnej lub szkoleniach*

*Identyfikacja wzorców i trendów w danych dotyczących zatrudnienia, które mogą wskazywać na potrzebę zmian w polityce personalnej lub szkoleniach*

Graphic: Jamie Champagne

## BI odpowiada na potrzeby biznesowe

Prognozy organizacyjne i operacyjne, które determinują lub nakierowują całe strategie biznesowe



*Analiza danych finansowych i operacyjnych, aby zidentyfikować obszary do optymalizacji kosztów i zwiększenia efektywności*

*Wykorzystanie analityki predykcyjnej do prognozowania trendów rynkowych i adaptacji strategii biznesowej*

*Wizualizacja danych i dashboardy zapewniające szybki dostęp do kluczowych informacji potrzebnych do podejmowania decyzji*

*Graphic: Jamie Champagne*

## BI odpowiada na potrzeby biznesowe

Oferta produktów biznesowych, terminy i rynki



*Analiza danych transakcyjnych i zachowań użytkowników w celu wykrywania potencjalnych oszustw lub nieprawidłowości*

*Monitorowanie wskaźników ryzyka i zgodności z przepisami, aby zapewnić przestrzeganie norm prawnych i branżowych*

*Ocena ryzyka kredytowego klientów i partnerów biznesowych, co pozwala na lepsze zarządzanie portfelem kredytowym*

*Graphic: Jamie Champagne*

Zwrot z Inwestycji (ROI) w Business Intelligence (BI), może obejmować (przykłady):



Identyfikacja rentownych segmentów klientów (Eliminacja relacji, które są zbyt kosztowne)



Prognozowanie finansów - Poprawa alokacji zasobów i efektywności operacyjnych



Analizowanie łańcuchów dostaw - prognozowanie zapasów, aby zmniejszyć koszty magazynowania



Zrozumienie zachowań klientów i rynków (inwestuj w technologię i/lub infrastrukturę tylko tam, gdzie są Twoi klienci i gdzie planują być

Graphic: Jamie Champagne

Zwrot z Inwestycji (ROI) w Business Intelligence (BI), może obejmować (powody):



1. Decyzje oparte na danych
2. Zwiększoną rentowność
3. Ulepszone budżetowanie i prognozowanie
4. Zwiększone zrozumienie klientów
5. Postępy technologiczne
6. BI typu self-service
7. Zrównoważoność inwestycji



Graphic: Jamie Champagne

Zwrot z Inwestycji (ROI) w Business Intelligence (BI), może obejmować (powody):



### 1. Decyzje oparte na danych:



*Firmy, niezależnie od ich głównego obszaru działalności - czy to finanse, marketing, czy obsługa klienta - coraz częściej opierają swoje operacje na danych. Narzędzia BI umożliwiają organizacjom efektywne wykorzystanie ich danych, przekształcając surowe dane w działania.*



Graphic: Jamie Champagne

Zwrot z Inwestycji (ROI) w Business Intelligence (BI), może obejmować (powody):



**1. Decyzje oparte na danych**

**2. Zwiększoną rentowność:**

*Dzięki identyfikacji najbardziej rentownych segmentów klientów, przedsiębiorstwa mogą efektywniej alokować swoje zasoby, koncentrując się na obszarach o najwyższych zwrotach. Takie ukierunkowane podejście pomaga eliminować segmenty nierentowne, zwiększając ogólną rentowność*



Graphic: Jamie Champagne

Zwrot z Inwestycji (ROI) w Business Intelligence (BI), może obejmować (powody):



**1. Decyzje oparte na danych**

**2. Zwiększoną rentowność**

**3. Ulepszone budżetowanie i prognozowanie:**

*Narzędzia BI zapewniają dokładniejsze możliwości prognozowania. Ta precyzaja pozwala na lepsze zarządzanie budżetem, zapewniając, że kapitał i zasoby są dostępne i wykorzystywane optymalnie w odpowiednich momentach*



Zwrot z Inwestycji (ROI) w Business Intelligence (BI), może obejmować (powody):



- 1. Decyzje oparte na danych**
- 2. Zwiększoną rentowność**
- 3. Ulepszone budżetowanie i prognozowanie:**
- 4. Zwiększone zrozumienie klientów**



*Integracja zewnętrznych danych o klientach z wewnętrznymi danymi biznesowymi za pomocą narzędzi BI pomaga w podejmowaniu mądrzejszych decyzji dotyczących oferty produktów i usług, zapewniając, że przedsiębiorstwa są zgodne z przyszłymi potrzebami klientów.*



Zwrot z Inwestycji (ROI) w Business Intelligence (BI), może obejmować (powody):



**1. Decyzje oparte na danych**

**2. Zwiększoną rentowność**



**3. Ulepszone budżetowanie i prognozowanie:**



**4. Zwiększone zrozumienie klientów**



**5. Postępy technologiczne:** Ciągła ewolucja big data, uczenia maszynowego i technologii sztucznej inteligencji rozszerzyła możliwości narzędzi BI, czyniąc je bardziej potężnymi i niezbędnymi dla organizacji

Graphic: Jamie Champagne

Zwrot z Inwestycji (ROI) w Business Intelligence (BI), może obejmować (powody):



**1. Decyzje oparte na danych**

**2. Zwiększoną rentowność**



**3. Ulepszone budżetowanie i prognozowanie:**



**4. Zwiększone zrozumienie klientów**



**5. Postępy technologiczne**

**6. BI typu self-service:** *Popyt na narzędzia BI typu self-service, które pozwalają użytkownikom w organizacji na dostęp i analizę danych bez specjalistycznych umiejętności technicznych, jest świadectwem rosnącej demokratyzacji danych. Ten trend podkreśla znaczenie BI w promowaniu kultury opartej na danych na wszystkich poziomach organizacji*

Graphic: Jamie Champagne

Zwrot z Inwestycji (ROI) w Business Intelligence (BI), może obejmować (powody):



- 1. Decyzje oparte na danych**
- 2. Zwiększoną rentowność**
- 3. Ulepszone budżetowanie i prognozowanie:**
- 4. Zwiększone zrozumienie klientów**
- 5. Postępy technologiczne**
- 6. BI typu self-service**
- 7. Zrównoważoność inwestycji:**



*Utrzymujący się zwrot z inicjatyw BI czyni je cenną inwestycją na przyszłość. Umiejętności i technologie rozwijane przez BI nie tylko mają zastosowanie dzisiaj, ale będą nadal relevantne i zapewniać przewagi konkurencyjne w nadchodzących latach*

Graphic: Jamie Champagne

**Program / narzędzia BI mogą być wykorzystywane do przekształcania danych w użyteczne informacje, które wspierają podejmowanie decyzji na wszystkich poziomach organizacji, od operacyjnego po strategiczny.**

*Graphic: Jamie Champagne*

## Ćwiczenie (0):

Jak dowiedzieć się, kto będzie subskrybować usługę ?

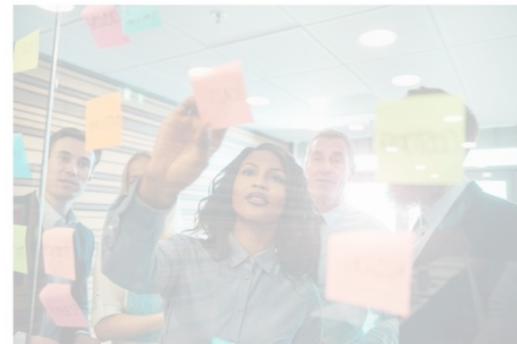
Przećwicz w wybranym narzędziu i przedstaw metodykę oraz wnioski (dyskusja, prezentacja wyników).

Dane: kanał Teams

Graphic: Jamie Champagne

Data Source: <https://archive.ics.uci.edu/dataset/222/bank+marketing>

# Cykl życia danych (DLC - DLM) a BI



**Business Goals**  
Defining business metrics and key performance indicators (KPIs)

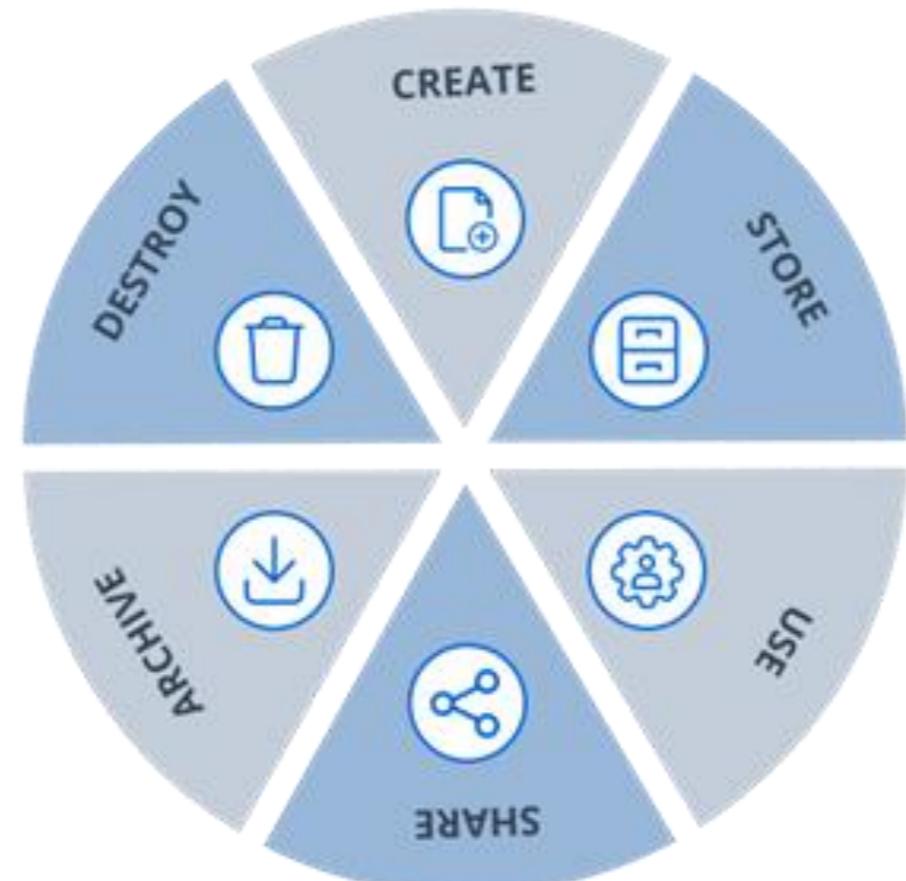


**Data Insights**  
**Data tools, technologies, and structure and governance in place to leverage the data**

- Polyzotis, Neoklis, et al. "Data lifecycle challenges in production machine learning: a survey." *ACM SIGMOD Record* 47.2 (2018): 17-28.
- D. C. Nguyen et al., "Enabling AI in Future Wireless Networks: A Data Life Cycle Perspective," in *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 553-595, Firstquarter 2021, doi: 10.1109/COMST.2020.3024783. keywords: {
- X. Yu and Q. Wen, "A View about Cloud Data Security from Data Life Cycle," *2010 International Conference on Computational Intelligence and Software Engineering*, Wuhan, China, 2010, pp. 1-4, doi:

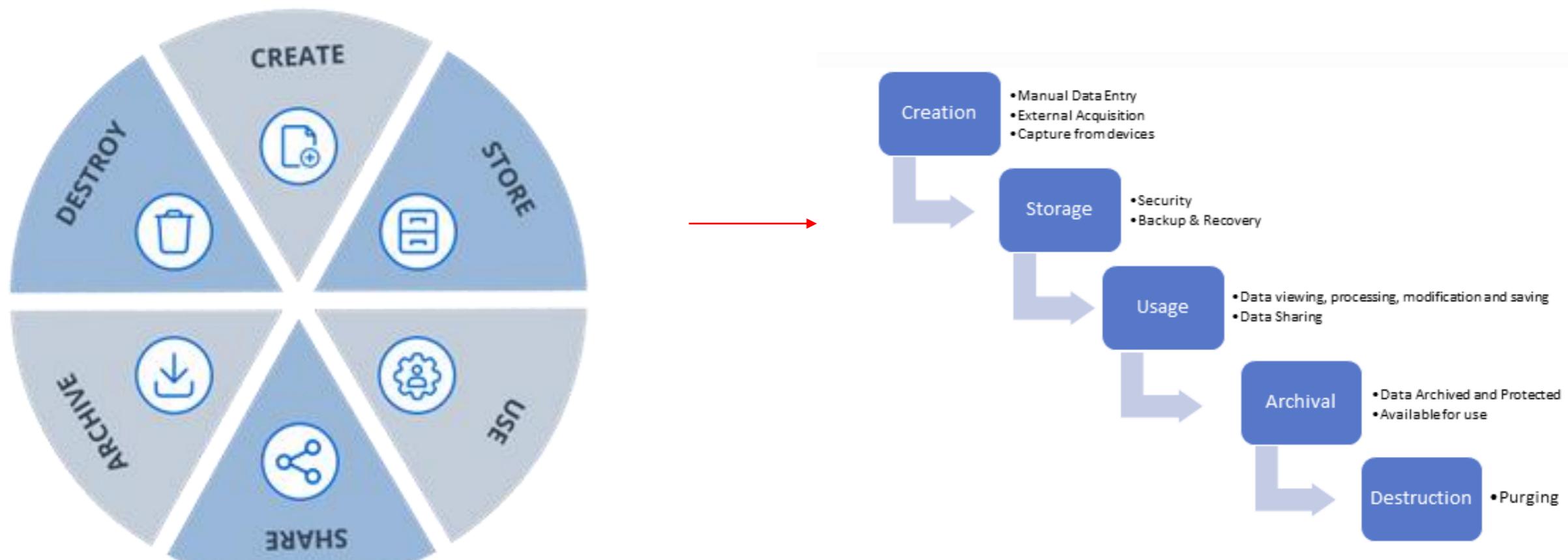
Graphic: Jamie Champagne

**Cykl życia danych** to etapy, przez które przechodzą dane od momentu ich utworzenia do momentu usunięcia (lub nie).



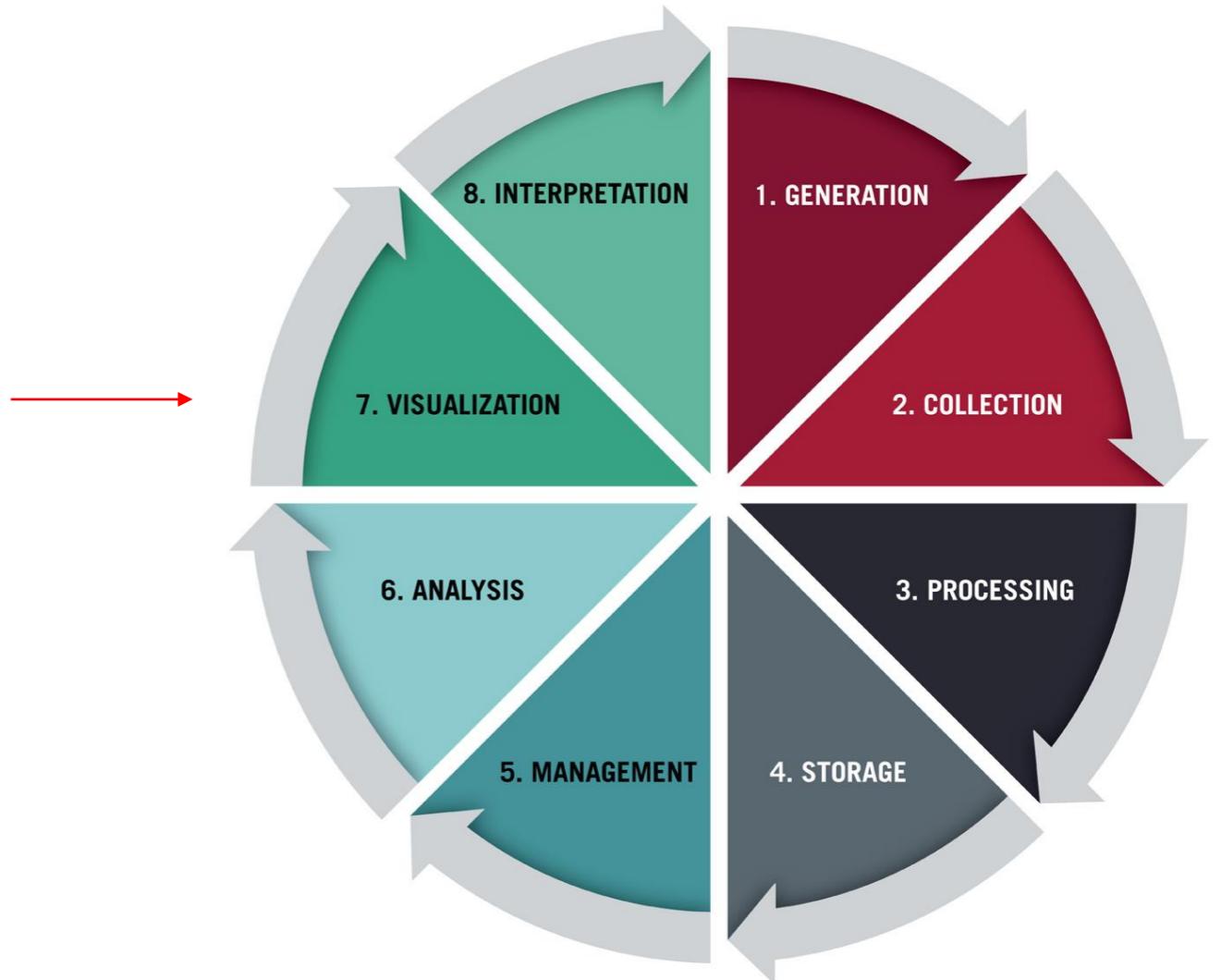
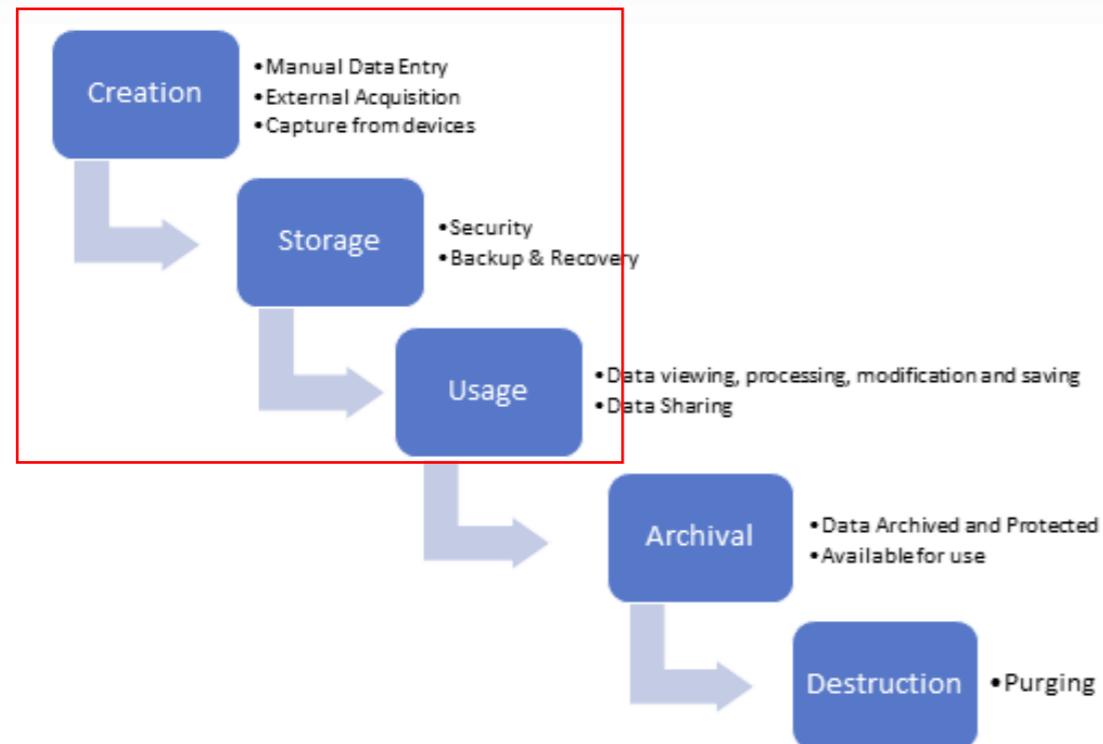
Graphics source: [www.dataworks.ie](http://www.dataworks.ie), [www.bakotech.pl](http://www.bakotech.pl)

**Cykl życia danych** to etapy, przez które przechodzą dane od momentu ich utworzenia do momentu usunięcia (lub nie).



Graphics source: [www.dataworks.ie](http://www.dataworks.ie), [www.bakotech.pl](http://www.bakotech.pl)

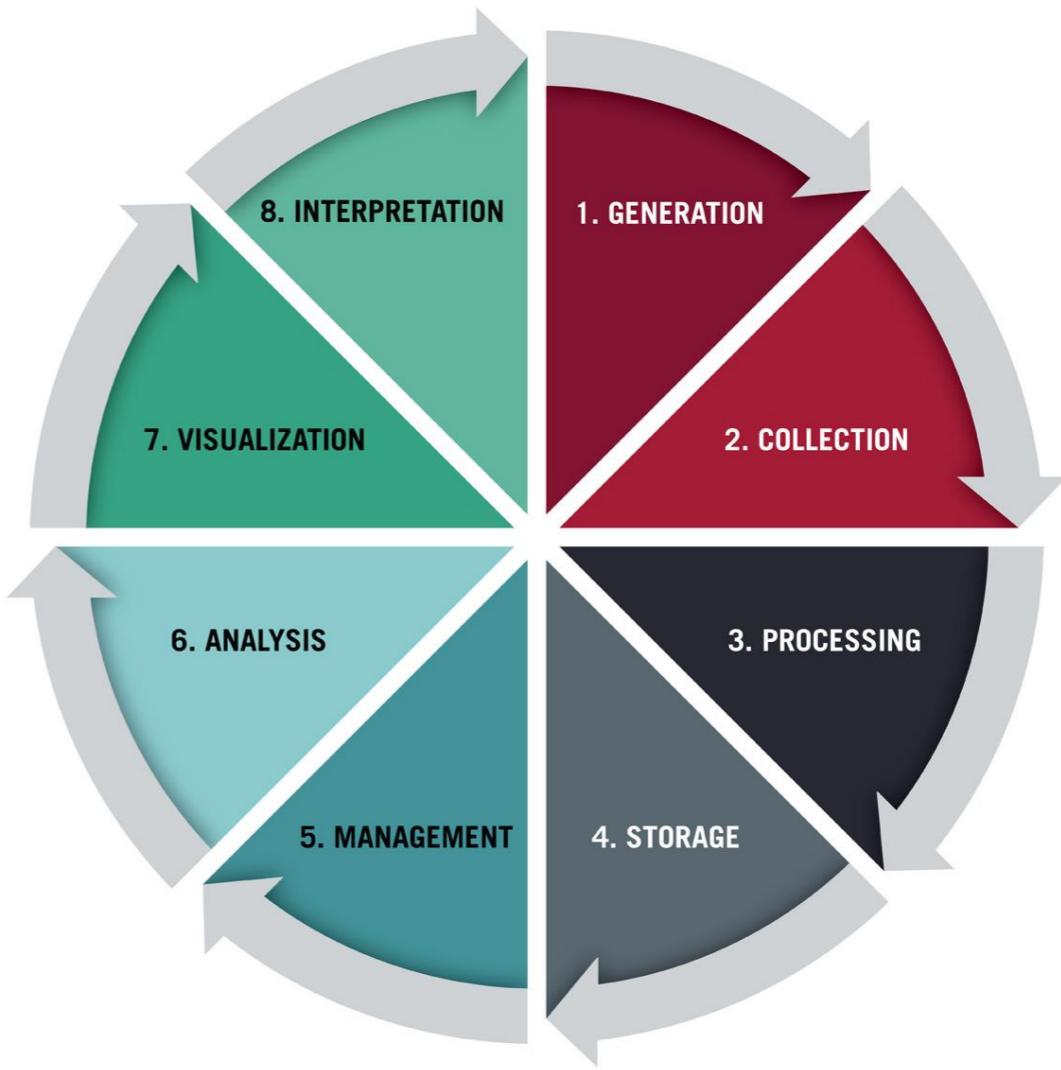
**Cykl życia danych** to etapy, przez które przechodzą dane od momentu ich utworzenia do momentu usunięcia (lub nie).



Graphic source: <https://segment.com/blog/data-life-cycle/>

# Generowanie danych

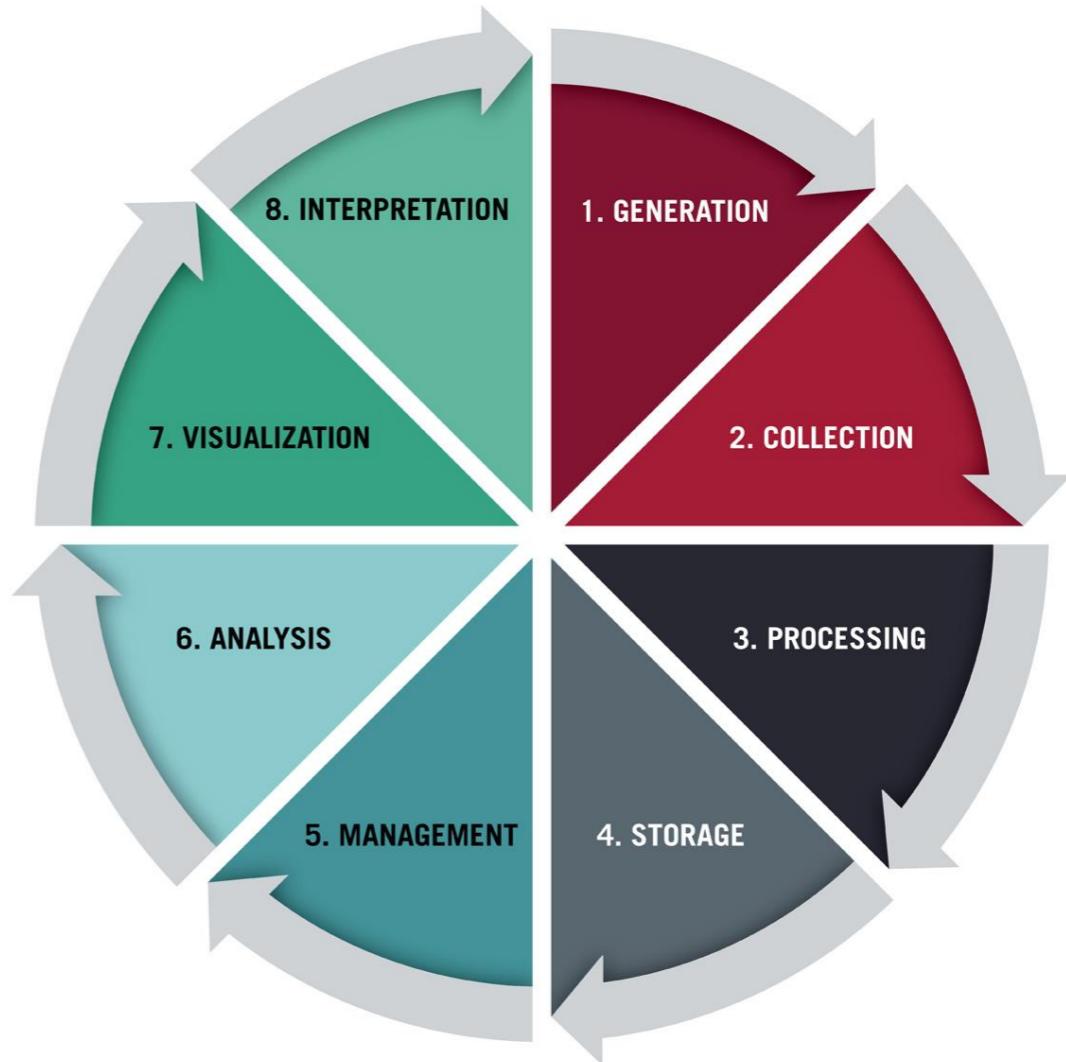
Proces tworzenia nowych danych



Graphic source: <https://segment.com/blog/data-life-cycle/>

## Zbieranie danych

Proces gromadzenia danych z istniejących źródeł



Graphic source: <https://segment.com/blog/data-life-cycle/>

## Generowanie danych

### Proces tworzenia nowych danych

Na tym etapie dane mogą być generowane zarówno wewnętrznie, jak i zewnętrznie, przez pracowników, klientów, dział sprzedaży, dział zasobów ludzkich lub poprzez inne formy komunikacji.

Każda aktywność przedsiębiorstwa jest źródłem generowania danych.

- **Ręczne tworzenie danych:** dane są wprowadzane ręcznie do pamięci (np. logowanie czasu pracy).
- **Automatyczne tworzenie danych:** dane są generowane automatycznie przez program komputerowy (np. czas przetwarzania zadania pracownika).
- **Symulowanie danych:** dane są symulowane, aby odzwierciedlać rzeczywiste dane (np. w ramach realizacji PoC).
- **Augmentacja danych:** istniejące dane są modyfikowane, aby stworzyć nowe dane (np. dodawanie subiektywnych informacji nt. rentowności, modyfikacja próbek materiałowych).

## Zbieranie danych

### Proces gromadzenia danych z istniejących źródeł

Proces akwizycji danych obejmuje różnorodne metody, takie jak komunikacja, obserwacja, przeprowadzanie wywiadów, dystrybucja ankiet, stosowanie formularzy oraz wykorzystanie urządzeń elektronicznych

- **Pozyskiwanie danych z istniejących źródeł:** dane są pobierane z istniejących źródeł, m.in. z systemów transakcyjnych, czujników, mediów społecznościowych i stron internetowych (np. API, Webscrapping)
- **Systemy CRM (Customer Relationship Management)** - gromadzą dane o klientach, kontaktach, potencjalnych klientach i szansach sprzedażowych
- **Systemy ERP (Enterprise Resource Planning) – Planowanie zasobów przedsiębiorstwa**
- **Analiza danych transakcyjnych** - umożliwia identyfikację trendów sprzedażowych i wzorców zakupowych
- **Obserwacja:** zbieranie wywiadu, ankiet
  - Obserwacja uczestnicząca: aktor staje się częścią obserwowanej grupy.
  - Obserwacja nieuczestnicząca: aktor obserwuje z zewnątrz.
- **Badanie dokumentów:** Analiza istniejących dokumentów, takich jak raporty, artykuły i książki (np. wykorzystanie OCR)
- **Social media:** Analiza postów, komentarzy i innych treści w mediach społecznościowych.
- **Czujniki i urządzenia IoT:** Automatyczne gromadzenie danych o środowisku i użytkowaniu.

# Ćwiczenie I – zbieranie danych

Pracujesz dla firmy zajmującej się sprzedażą produktów. Organizacja chce wprowadzić nową usługę – sprzedaży książek.

Twoja organizacja, stawia sobie za cel, opracowanie planu tego przedsięwzięcia, którego jednym z założeń, jest analiza rynku.

Prosi Cię o cykliczne dostarczanie informacji **na temat aktualnie sprzedawanych książek przez konkurencję** (możliwie jak najwięcej informacji – ale nie wie jeszcze, które będą potrzebne).

Twoim zadaniem, jest przygotowanie narzędzia, aby pobierać cyklicznie takie dane:

- Na temat sprzedawanych książek (np. ceny, tytuły) – aby twoja organizacja mogła opracować własny cennik
- Na temat konkurencji – aby mogła ustalić, jak monitorować zachowania konkurencyjne.

Wykonanie: 13.03 (zajęcia lub indywidualnie), 20.03 (zajęcia lub indywidualnie), Wnioski do: 27.03.2024 (mail)

# Ćwiczenie 1 – zbieranie danych (webscraping)

1. Twoim zadaniem będzie pobieranie informacji na temat książek (dostęp do tytułów, cen, możliwe jak najwięcej informacji), aby można było monitorować w przyszłości trendy finansowe. Przygotuj prosty skrypt (np. w Pythonie `requests`, `BeautifulSoup`), który będzie pobierał dane o książkach z przykładowej strony internetowej zajmującej się recenzowaniem lub sprzedażą książek (min. np. Books to Scrape).

## Kroki do wykonania:

1. Zapoznanie z narzędziami - Przegląd dokumentacji obu bibliotek
2. Analiza struktury strony: odwiedzić wybraną stronę (upewnij się ze scrapowanie jest dozwolone: `robot.txt`) i za pomocą narzędzi deweloperskich przeglądarki przeanalizuj strukturę strony, w szczególności jak zorganizowane są informacje o książkach)
3. Napisanie skryptu: Zimportuj niezbędne biblioteki, Napisz funkcję wysyłającą żądania do strony internetowej, dokonaj parsowania zawartości HTML, wyodrębnianie danych
4. Po zebraniu wszystkich danych zapisz je do pliku CSV lub innego wybranego formatu (np. do DataFrame i -> CSV).
5. Zadbaj o paginacje (jeśli wybrana strona, posiada więcej stron z produktami)
6. Dodaj obsługę błędów
7. Upewnij się, że skrypt zawiera przerwy między żądaniami, aby uniknąć przeciążenia serwera, np. używając `time.sleep()`

# Ćwiczenie 2 – zbieranie danych (API)

2. Twoim zadaniem będzie weryfikacja, czy możliwe jest pobranie danych (**jeśli tak, to jakich ?**) o konkurencji ze źródeł Głównego Urzędu statystycznego:

API DBW:

<https://api.stat.gov.pl/Home/DBWApi>

<https://api-dbw.stat.gov.pl/apidocs/index.html>

Dane udostępniane są poprzez REST-owe API w formacie JSON, CSV

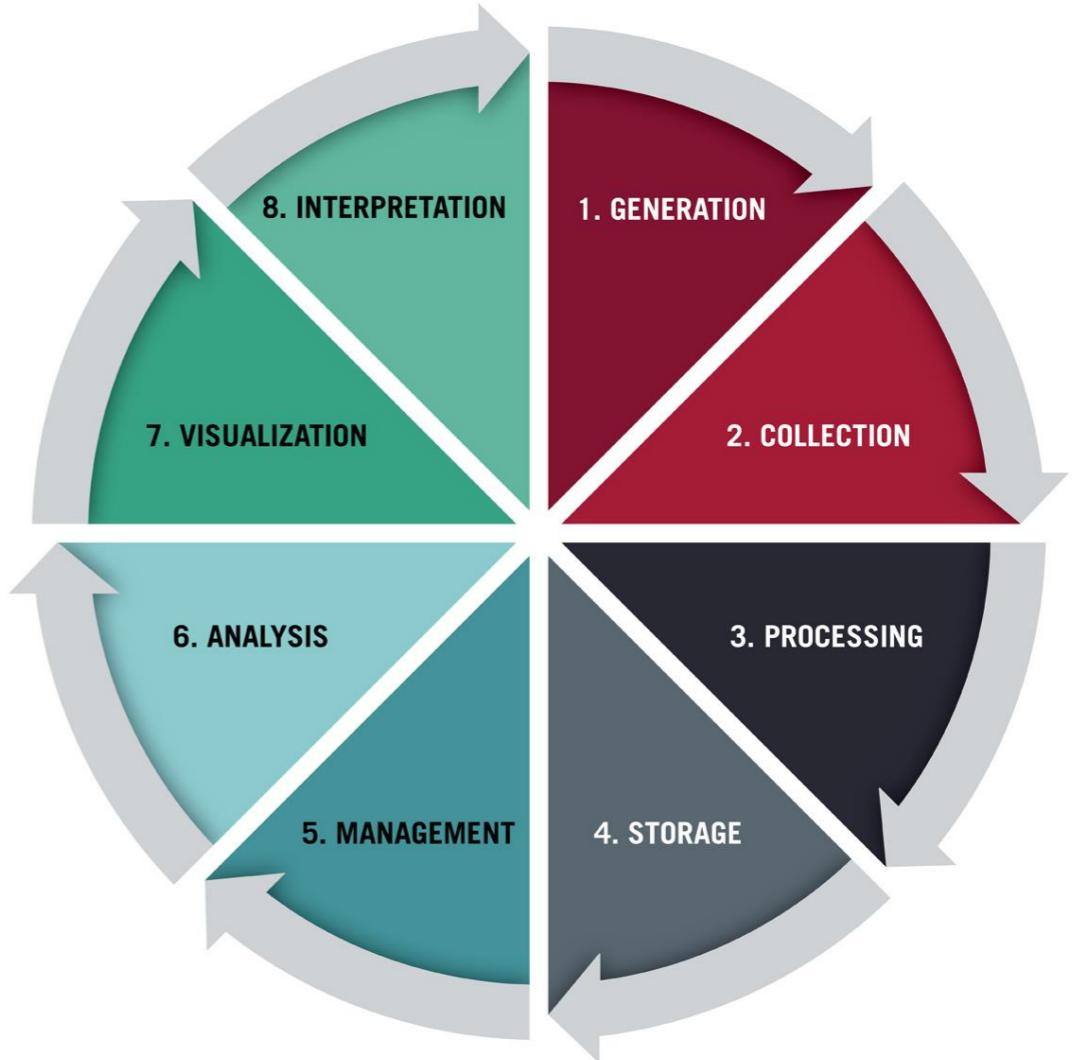
Anonimowy użytkownik (niezalogowany) może wykonać do 5 żądań na sekundę, 100 żądań na 15 minut, 1000 żądań na 12 godzin, 10000 żądań na 7 dni

Zarejestrowany użytkownik może wykonać do 10 żądań na sekundę, 500 żądań na 15 minut, 5000 na 12 godzin, 50000 na 7 dni

Przygotuj opis wykonanych prac (sprawozdanie) i przedstaw prowadzącemu. Jeśli wykonujesz prace zdalnie, prześlij prowadzącemu sprawozdanie z opisem wykonanych czynności.

▼ 77:	
id-przekroj:	736
id-wymiar:	563
nazwa-wymiar:	"Towary i usługi konsumpcyjne_1"
id-pozycja:	6656166
▼ nazwa-pozycja:	"czasopisma, gazety, książki oraz artykuły piśmienne, kreślarskie, malarstkie"

## Przetwarzanie i magazynowanie danych



Graphic source: <https://segment.com/blog/data-life-cycle/>

## Przetwarzanie danych

Dzieje się tak, gdy oprogramowanie pobiera surowe dane i przekształca je w coś użytecznego. Metodyki czyszczą, korygują, kompresują, szyfrują i tłumaczą dane.

## Przechowywanie danych

Należy przechowywać dane, gdy już są w organizacji. W tym celu trzeba chronić je przed włamaniami i utratą, tworząc kopie zapasowe danych, niezależnie od tego, czy przechowywane są na miejscu, czy w chmurze.

# Przetwarzanie danych

**Analiza statystyczna:** Podstawowa metoda przetwarzania danych, obejmująca obliczanie statystyk opisowych, takich jak średnia, mediana, odchylenie standardowe, oraz przeprowadzanie testów statystycznych do weryfikacji hipotez.

**Przetwarzanie języka naturalnego (NLP):** Techniki wykorzystywane do analizy, rozumienia i interpretacji ludzkiego języka, wykorzystywane w takich aplikacjach jak analiza sentymentu, tłumaczenie maszynowe i chatboty.

**Wizualizacja danych:** Tworzenie graficznych przedstawień danych, które pomagają w ich zrozumieniu i interpretacji, wykorzystując narzędzia takie jak wykresy, grafy, mapy ciepła i inne.

**Uczenie maszynowe:** Zastosowanie algorytmów, które uczą się z danych i mogą przewidywać wyniki lub kategoryzować dane na podstawie wcześniejszych obserwacji. Wśród popularnych technik znajdują się algorytmy klasyfikacji, regresji, klasteryzacji i sieci neuronowe.

**Przetwarzanie sygnałów:** Analiza danych w formie sygnałów, takich jak dźwięk, obrazy czy sygnały biomedyczne, wykorzystując techniki takie jak transformata Fouriera, filtracja cyfrowa czy analiza składowych głównych (PCA). Uwaga: obrazy, dźwięki to także sygnały.

**Data Mining:** Proces odkrywania wzorców i zależności w dużych zbiorach danych poprzez stosowanie metod takich jak klasteryzacja, detekcja anomalii i reguły asocjacyjne.

**Eksploracyjna analiza danych (EDA):** Proces analizy zbiorów danych w celu podsumowania ich głównych cech, często przy użyciu wizualizacji, przed dokonaniem bardziej formalnej analizy statystycznej.

# Magazynowanie danych

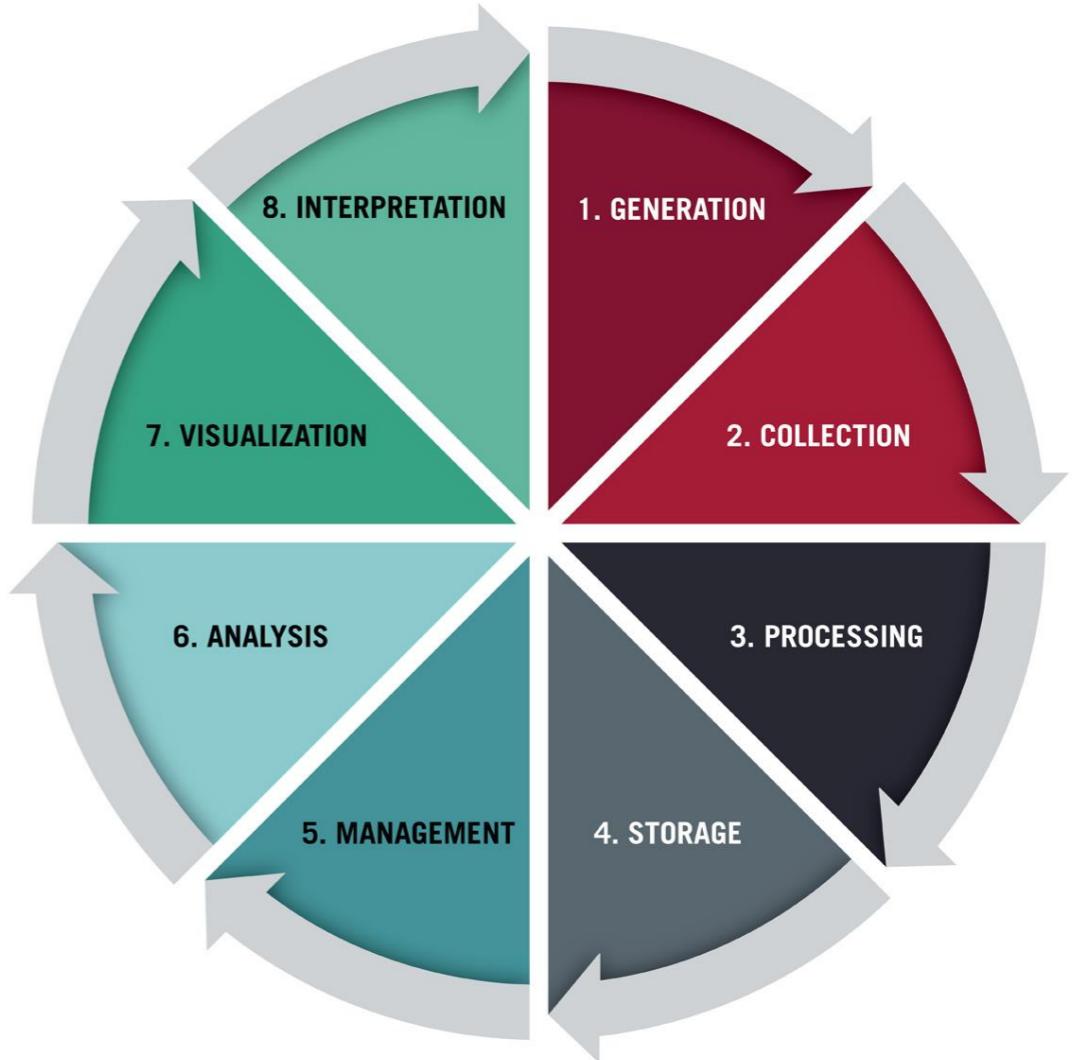
- **Lokalna pamięć masowa:**
  - Dyski twarde (HDD)
  - Dyski SSD (Solid State Drive)
  - Pamięć masowa NAS (Network Attached Storage)
- **Pamięć masowa w chmurze:**
  - Publiczna:
    - Infrastruktura chmurowa jest udostępniana wielu organizacjom.
    - Dostępna jest przez Internet.
    - Oferuje łatwość użytkowania i skalowalność.
    - Może być tańsza niż chmura prywatna
  - Prywatna:
    - Infrastruktura chmurowa jest dedykowana dla jednej organizacji i nie jest udostępniana innym.
    - Może być zlokalizowana w centrum danych organizacji lub u zewnętrznego dostawcy usług hostingowych.
    - Oferuje wysoki poziom kontroli i bezpieczeństwa.
    - Może być droższa niż chmura publiczna.

# Magazynowanie danych

- **Pamięć masowa w chmurze cd.:**
  - Publiczna, np...:
    - AWS - Amazon S3, EMR, Athena,
    - Microsoft Azure Blob Storage, Azure SQL Data Warehouse, Azure HDInsight, Azure Data Lake Store,
    - GCP (Google Cloud Platform, BigQuery, Cloud Storage, Cloud Dataproc, Cloud Dataflow)
    - Snowflake (DaaS)
    - IBM Cloud (IBM Cloud SQL Data Warehouse, Cloud Object Storage, IBM Cloud Big Data Platform, IBM Cloud Data Lake)
- **Systemy zarządzania bazami danych:**
  - Relacyjne bazy danych (np. MySQL, PostgreSQL, Oracle)
  - Bazy danych NoSQL (np. MongoDB, Cassandra)

## Zarządzanie danymi

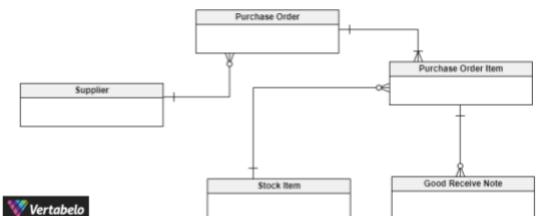
Zarządzanie pomaga organizować, przechowywać i pobierać dane



Graphic source: <https://segment.com/blog/data-life-cycle/>

## Zarządzanie danymi:

- **Modelowanie danych** – tworzenie abstrakcyjnych modeli, które reprezentują strukturę danych w organizacji, często przy użyciu diagramów ERD (Entity-Relationship Diagram). Proces tworzenia wizualnych schematów i planów, które reprezentują i organizują strukturę danych w systemie:
  - Model **koncepcyjny** - definiuje ogólną strukturę firmy i danych. Służy on do organizowania koncepcji biznesowych, zdefiniowanych przez interesariuszy firmy oraz architektów danych.
    - Diagramy ER (Entity-Relationship): Używane do przedstawienia związków między bytami w bazie danych. Są one prostymi diagramami pokazującymi byty jako prostokąty, a relacje jako linie łączące te prostokąty, z opcjonalnymi "skrzydłami" wskazującymi kardynalność (1:1, 1:n, n:m).
    - Schematy UML (Unified Modeling Language): Choć częściej stosowane w projektowaniu oprogramowania, mogą być również wykorzystywane do koncepcyjnego modelowania danych.
  - Model **logiczny** - opiera się na modelu koncepcyjnym z określonymi atrybutami danych w obrębie poszczególnych encji oraz relacjami między tymi atrybutami:
    - Schematy normalizacji: Proces normalizacji bazy danych prowadzi do stworzenia takiego modelu, który eliminuje redundancję i zależności funkcyjne. Normalizacja skupia się na rozkładaniu tabel na mniejsze, połączone relacje z zachowaniem integralności danych.
    - Model relacyjny: Używany do przedstawienia danych w postaci tabel, które są powiązane ze sobą przez klucze obce.
  - Model **fizyczny**: Fizyczny model danych to określona implementacja logicznego modelu danych. Jest on tworzony przez deweloperów i administratorów baz danych. Model tego typu jest opracowywany z myślą o wybranym narzędziu do obsługi baz danych, technologii magazynowania danych oraz konektorów danych, które umożliwiają zarządzanie danymi na platformach biznesowych w sposób wybrany przez użytkowników.



## Zarządzanie danymi cd:

- **Normalizacja** – proces projektowania schematu bazy danych w taki sposób, aby zminimalizować redundancję i zależności.
- **Zarządzanie metadanymi** – przechowywanie informacji o danych, takich jak ich struktura, format i zasady dotyczące ich użycia.
- **Bezpieczeństwo danych** – ochrona danych przed nieautoryzowanym dostępem, naruszeniem lub utratą poprzez szyfrowanie, kontrolę dostępu itp.
- **Backup i recovery** – tworzenie kopii zapasowych danych i ich odzyskiwanie w przypadku awarii.
- **Data governance** – zestaw polityk i procedur zarządzających dostępem, wykorzystaniem i jakością danych w organizacji.

Przykładowy proces modelowania danych:

1.Określenie wymagań:

*Rozmowa z interesariuszami i zbieranie wymagań biznesowych oraz zrozumienie, jakie informacje są zbierane, przechowywane i jak będą wykorzystywane.*

2.Tworzenie Modelu Konceptualnego:

*Opracowanie diagramu ER, który identyfikuje główne byty, ich atrybuty i relacje.*

3.Rozwinięcie Modelu Logicznego:

*Konwersja diagramu ER do modelu relacyjnego z tabelami, kluczami i relacjami.*

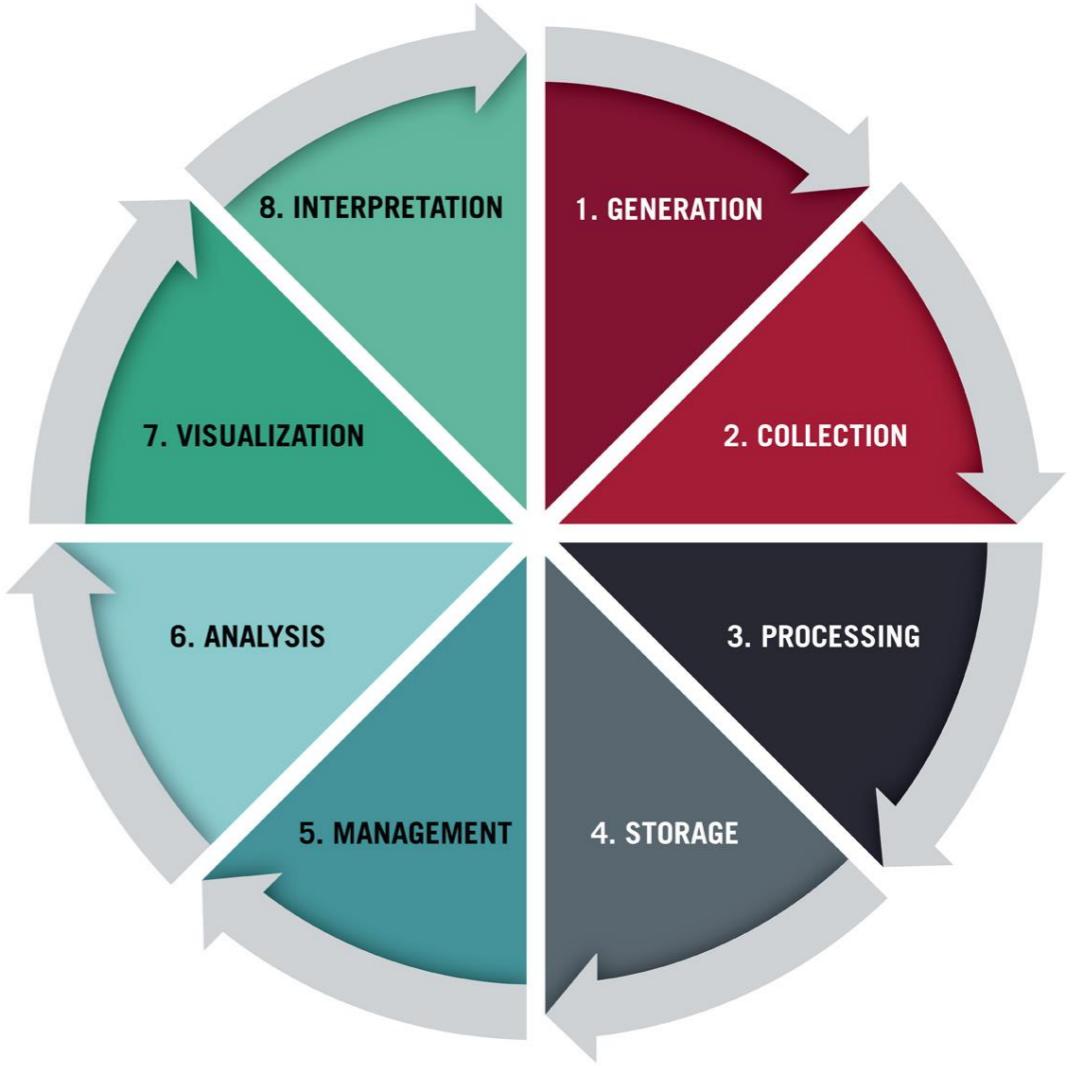
4.Implementacja Modelu Fizycznego:

*Przekształcenie modelu logicznego w schematy, które będą rzeczywiście implementowane w wybranej bazie danych.*

5.Weryfikacja i Optymalizacja:

*Testowanie modelu pod kątem wydajności i integralności danych oraz dokonywanie niezbędnych poprawek i optymalizacji.*

## Analiza danych



Graphic source: <https://segment.com/blog/data-life-cycle/>

Dla większości firm analiza danych jest najważniejszą częścią cyklu życia danych. Surowe dane, po przetworzeniu, używane są do późniejszej analizy. Proces przekształcania nieprzetworzonych danych w użyteczne informacje z wykorzystaniem matematycznych modeli:

- Statystyka opisowa - Umożliwia podsumowanie i opisanie cech zbiorów danych za pomocą miar takich jak średnia, mediana, moda, zakres, odchylenie standardowe, wariancja i kwartyle. Statystyka opisowa jest często punktem wyjścia w analizie danych.
- Regresja - Jest to metoda statystyczna używana do modelowania związku między zmienną zależną a jedną lub więcej zmiennymi niezależnymi. Najpopularniejsze to regresja liniowa i regresja logistyczna. Umożliwiają przewidywanie wartości zmiennej zależnej na podstawie znanych wartości zmiennych niezależnych.
- Analiza wariancji (ANOVA) - Pozwala porównać średnie między różnymi grupami, aby stwierdzić, czy istnieją statystycznie istotne różnice między nimi. Jest szczególnie przydatna, gdy badamy wpływ jednej lub więcej zmiennych kategorycznych na zmienną ciągłą.
- Analiza skupień (clustering) - Jest metodą uczenia nienadzorowanego, która grupuje podobne do siebie obiekty w danej przestrzeni. Stosowana jest do identyfikacji naturalnych podziałów w zbiorze danych, bez wcześniejszego definiowania kategorii.
- Analiza głównych składowych (PCA) - Technika redukcji wymiarowości, która przekształca zbiór możliwie skorelowanych zmiennych w zestaw wartości nieskorelowanych zwanych głównymi składowymi. PCA jest często stosowane do uproszczenia danych przed dalszą analizą.
- Sieci neuronowe i głębokie uczenie - Metody te, oparte na sztucznej inteligencji, są stosowane do modelowania złożonych wzorców i zależności w danych. Sieci neuronowe są szczególnie przydatne w rozpoznawaniu obrazów, przetwarzaniu języka naturalnego i innych zadań wymagających analizy dużych zbiorów danych.
- Maszyny wektorów nośnych (SVM) - Metoda uczenia maszynowego używana do klasyfikacji i regresji. SVM działa poprzez znajdowanie hiperpłaszczyzny w przestrzeni wielowymiarowej, która najlepiej oddziela różne klasy danych.
- Drzewa decyzyjne i lasy losowe - Te metody są używane do klasyfikacji i regresji. Tworzą model, który przewiduje wartość zmiennej przez naukę prostych reguł decyzyjnych wywnioskowanych z danych.
- I wiele innych ...

# Ćwiczenie 3 – analiza subskrybcji

Plik „Ocena subskrypcji.csv” (Teams) zawiera takie informacje jak wiek (age), zawód (job), stan cywilny (marital), edukacja (education), i inne, aż do decyzji o subskrypcji (y).

Zbadaj:

- jak poszczególne cechy (zmienne niezależne) wpływają na decyzję o subskrypcji (zmienna zależna y). Przeprowadź analizę statystyczną, aby ocenić rozkłady parametrów, np. :
  - Jaki jest wiek (średni) klientów
  - Jaki jest najczęstszy zawód
  - Ile procent ankietowanych zdecydowało się na zakup subskrypcji
  - Sprawdź korelację
  - Spróbuj przewidzieć, czy klient zdecyduje się na subskrypcję (podaj dokładności wykorzystanych metod). Aby przewidzieć, czy klient zdecyduje się na subskrypcję, można zastosować model klasyfikacyjny, spróbuj przetestować wybrane i nakreślić ich możliwości i ograniczenia (np. model wymaga danych numerycznych, posiadane dane są jakościowe. Możliwe rozwiązanie: kategoryzacja itp.): Regresja logistyczna, Drzewa decyzyjne, Las losowy, Maszyny wektorów nośnych (SVM), Sieci neuronowe

Wykonanie: 20.03 (zajęcia lub indywidualnie), 27.03 (zajęcia lub indywidualnie)

Wnioski do: 03.04.2024 (mail)

# Ćwiczenie 4 – analiza subskrybcji (regresja logistyczna)

Wykorzystaj dowolne narzędzie (Matlab, Python, Excel Dodatek Analizy danych, lub inny dowolny) i wykonaj regresję logistyczną, aby spróbować przewidzieć, czy klient zdecyduje się na subskrypcję. Regresja logistyczna jest popularnym modelem klasyfikacyjnym używanym w uczeniu maszynowym, który jest stosowany do przewidywania wyniku binarnego (tak/nie, 1/0) na podstawie jednej lub wielu zmiennych niezależnych.

Przed zastosowaniem regresji logistycznej, przygotuj dane:

- czy regresja logistyczna wymaga danych numerycznych ? Jeśli tak, zakodowanie zmiennych kategorycznych, takich jak job, marital, education, itp.
- Rozdzielenie danych na zbiór treningowy i testowy: Pozwoli to na ocenę wydajności modelu na danych, których nie widział podczas treningu.
- Normalizacja danych: czy regresja logistyczna wymaga normalizacji ? Sprawdź jaki wpływ ma normalizacja dla efektywności ? Czy może wpływać pozytywnie, np.. gdy zmienne mają różne zakresy wartości ?
- Przygotuj sprawozdanie i opisz metodykę, uzyskane wyniki i wnioski. Wnioski prześlij prowadzącemu, przedyskutuj wyniki.

Wykonanie: 20.03 (zajęcia lub indywidualnie), 27.03 (zajęcia lub indywidualnie)

Wnioski do: 03.04.2024 (mail)

# Ćwiczenie 5 – analiza subskrypcji (modele oparte na sieciach neuronowych)

Wykorzystaj dowolne narzędzie (Matlab, lub inny dowolny, np. scikit-learn: machine learning in Python) i wykonaj analizę w wykorzystaniem modeli opartych na sieciach neuronowych, aby spróbować przewidzieć, czy klient zdecyduje się na subskrypcję. Oceń, który model może zostać użyty (np. modelu sieci neuronowej z jedną lub kilkoma warstwami ukrytymi, MLPClassifier, Perceptron etc.), aby zobaczyć, jak radzi sobie z takim zadaniem klasyfikacyjnym.

- Czy niezbędne było kodowanie zmiennych kategorycznych, takich jak job, marital, education, itp.?
- Czy potrzebne było mapowanie zmiennej docelowej na wartości binarne
- Czy zastosowałeś technikę normalizacji ? Dlaczego ?
- Jak wyglądał podział zbioru treningowego i walidacyjnego ?

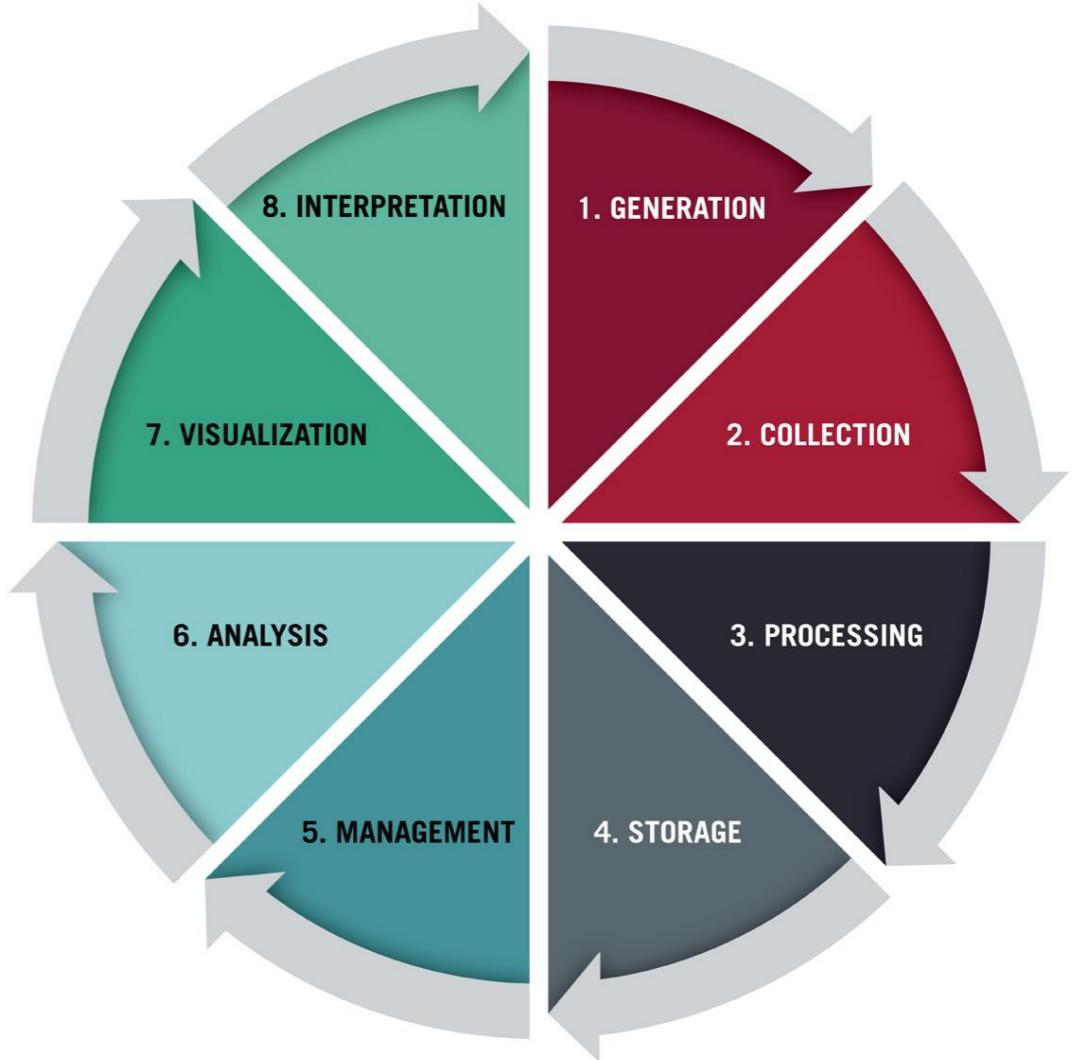
Przygotuj sprawozdanie i opisz metodykę, uzyskane wyniki i wnioski. Prześlij prowadzącemu, przedyskutuj wyniki.

Wykonanie: 20.03 (zajęcia lub indywidualnie), 27.03 (zajęcia lub indywidualnie)

Wnioski do: 03.04.2024 (mail)

## Wizualizacja danych

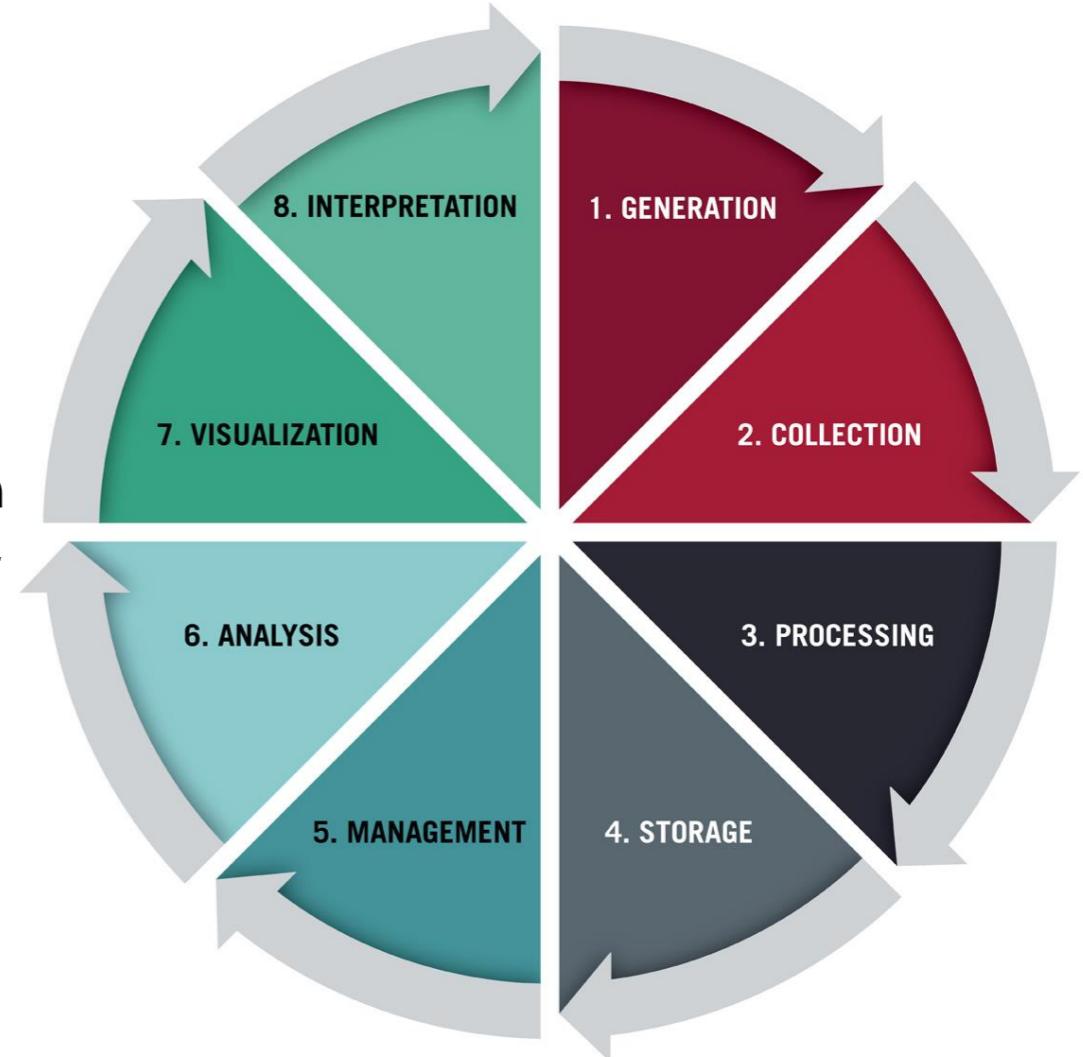
W tym miejscu oprogramowanie przekształca użyteczne dane w atrakcyjne wizualnie formaty, które personel może zobaczyć.



Graphic source: <https://segment.com/blog/data-life-cycle/>

## Interpretacja i publikacja danych

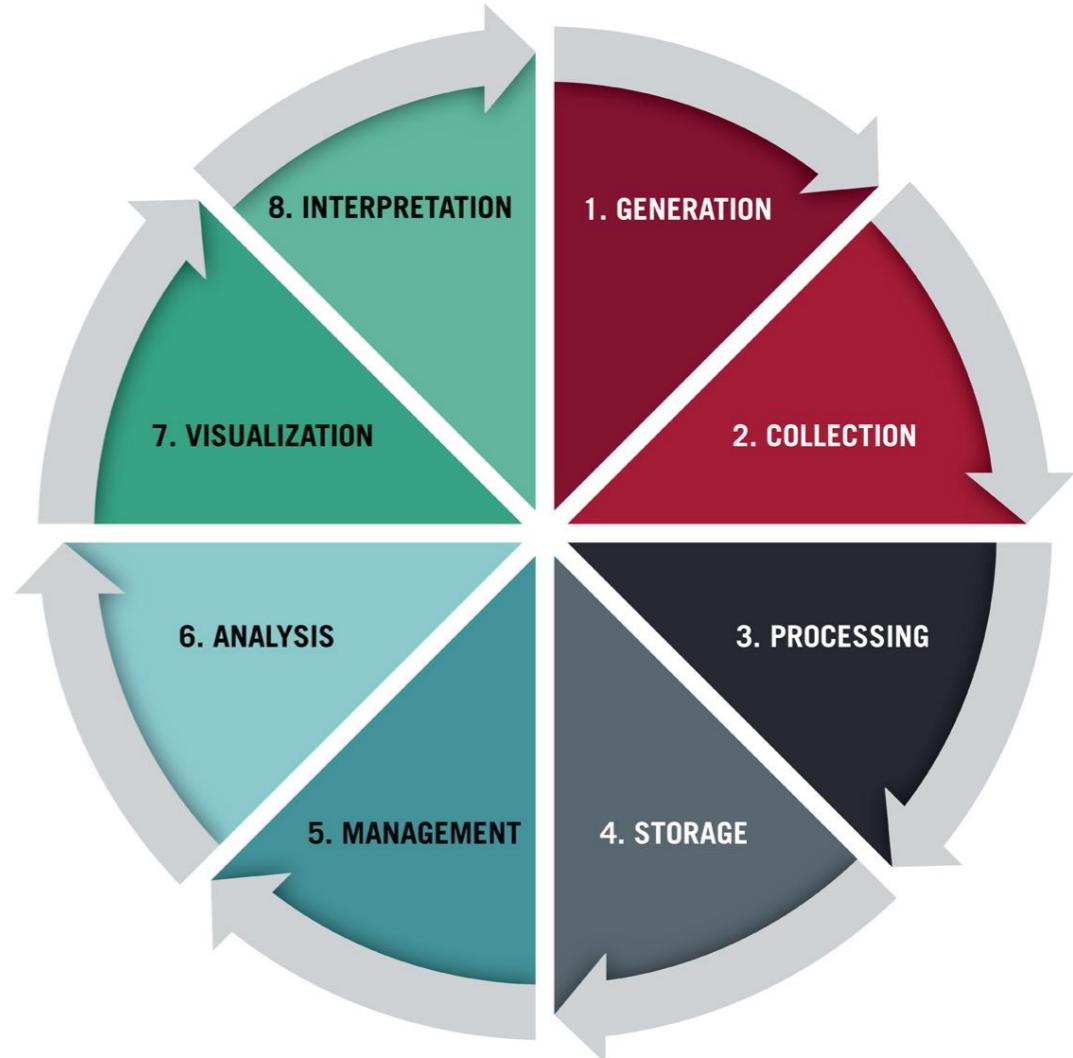
*Iteracyjne używanie danych i ich re-interpretacje w celu podjęcia decyzji.*



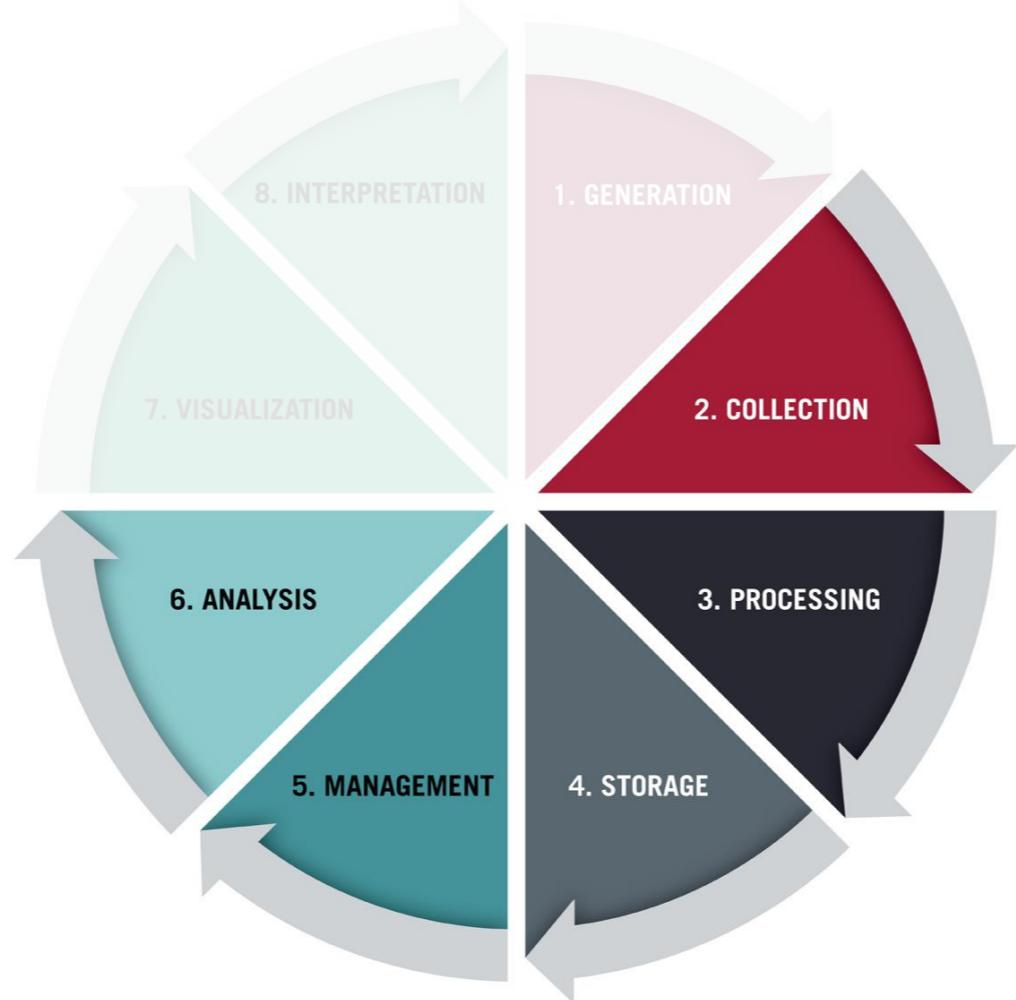
Graphic source: <https://segment.com/blog/data-life-cycle/>

## Opcjonalnie - Niszczenie danych

Jest to ostatni krok cyklu życia danych. Gdy dane przestaną być wykorzystywane, mogą zostać odpowiednio, profesjonalnie, zniszczone w 100%. Może zaistnieć potrzeba zniszczenia wrażliwych danych ze względu na kwestie prawne lub też związane ze zgodnością, lub koszt przechowywania stanie się zbyt uciążliwy.



Graphic source: <https://segment.com/blog/data-life-cycle/>



# Cykl jakości danych – inna perspektywa

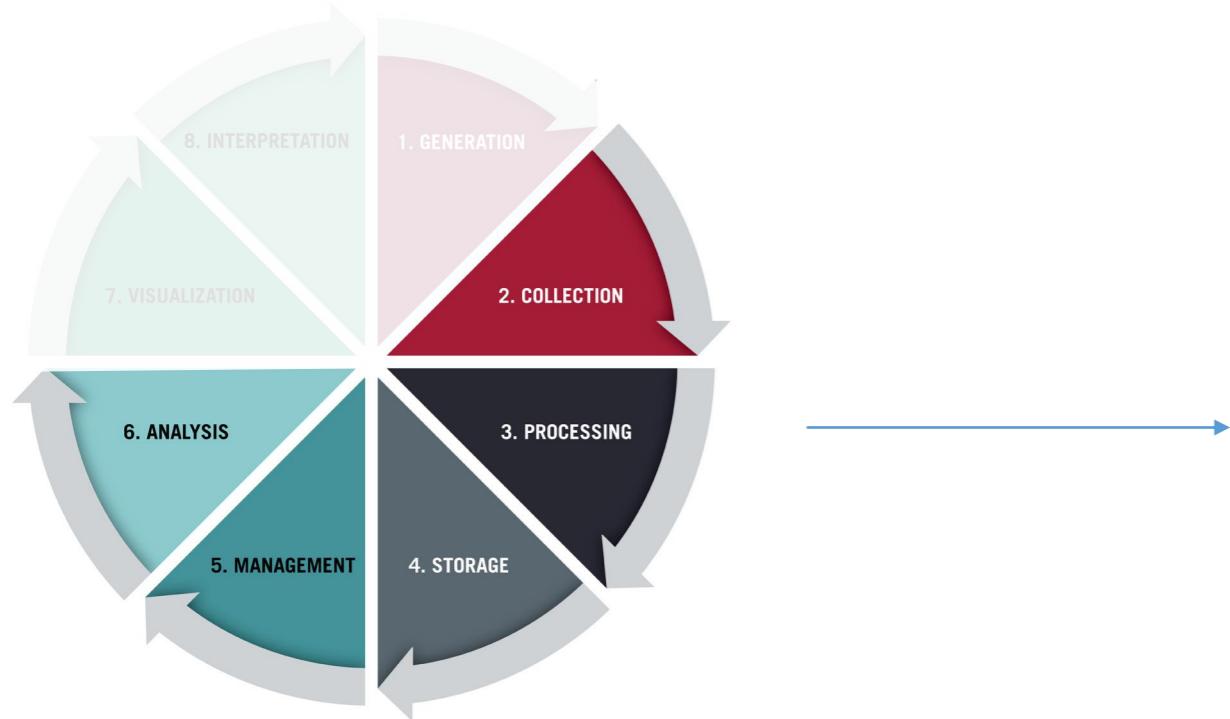
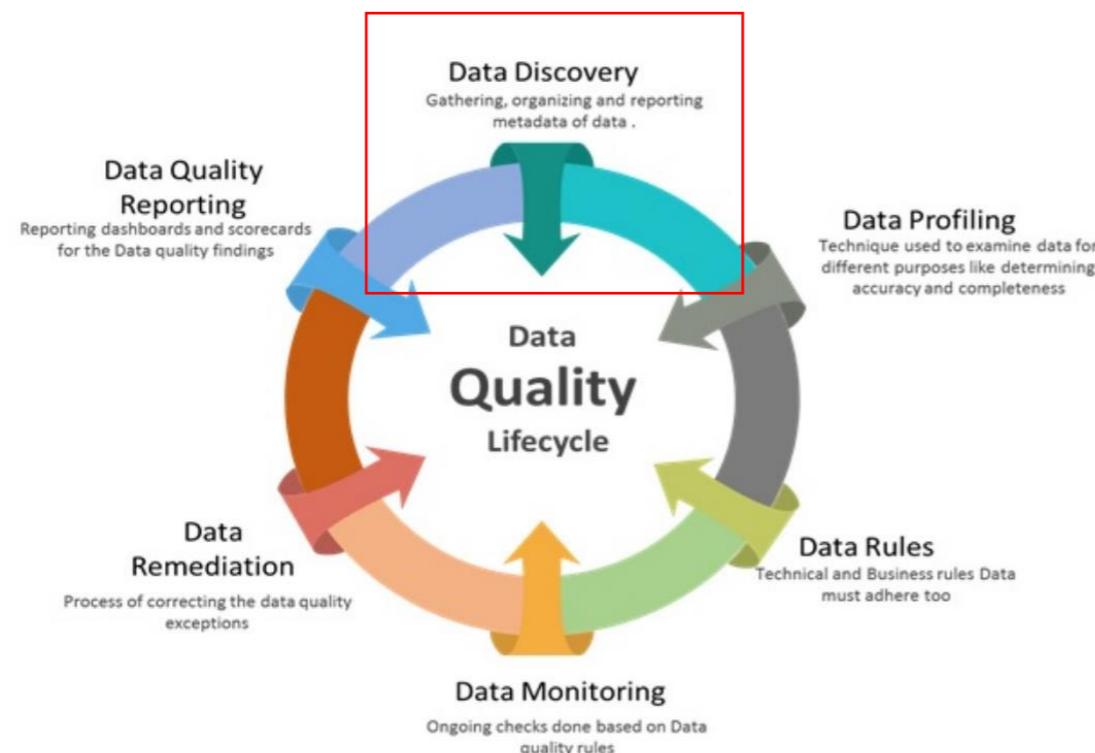


Image: [https://media.licdn.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image-shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU](https://media.licdn.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image-shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU)

# Cykl jakości danych – inna perspektywa



Odkrywanie danych (eksploracja danych lub odkrywanie wiedzy), odnosi się do procesu przeszukiwania i analizowania dużych zbiorów danych w celu odkrycia ukrytych wzorców, trendów i anomalii.

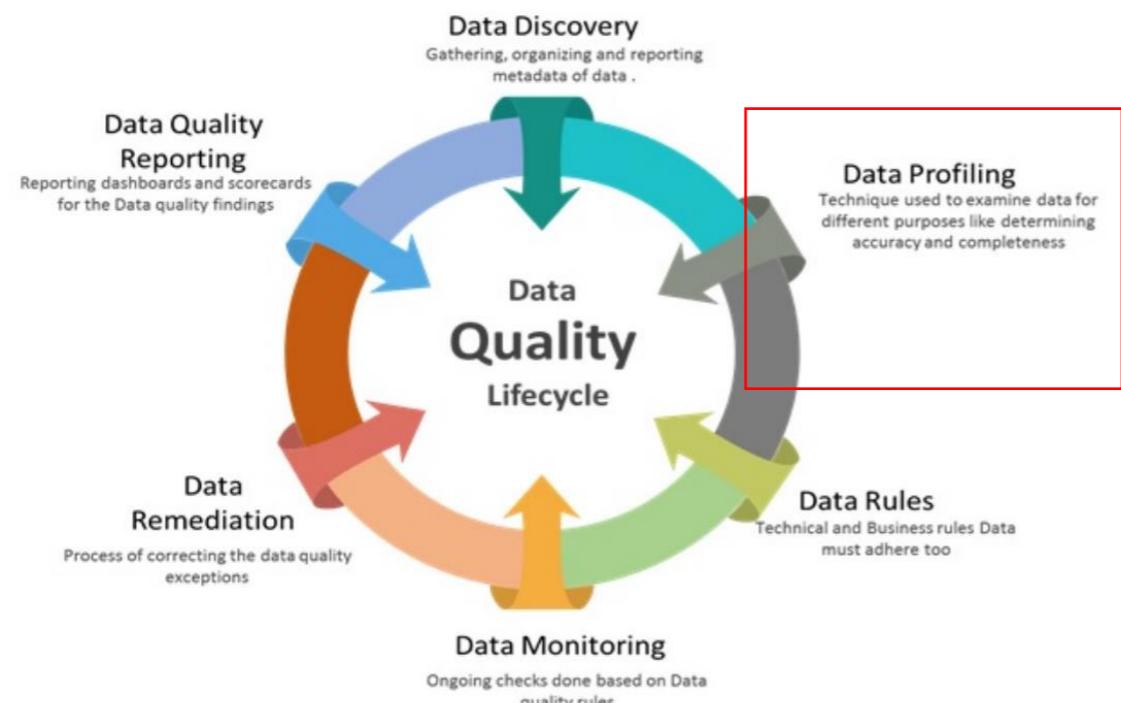
Techniki „data discovery”, odnoszą się do procesów i narzędzi używanych do identyfikacji, analizy i wizualizacji danych w celu uzyskania wglądów i informacji. Proces odkrywania danych jest kluczowym elementem analizy biznesowej, pomagając organizacjom lepiej zrozumieć swoje dane, odkrywać nowe wzorce, trendy i anomalie, a także wspierać podejmowanie decyzji.

Obejmuje metody eksplorowania zestawów danych (np. eksploracyjna EDA, Wizualizacje, Mining, Clusterization, ML, PCA etc.) w celu podsumowania ich głównych cech.

Korzystając z metod statystycznych i wizualizacji, możesz dowiedzieć się więcej o zestawie danych, aby określić jego gotowość do analizy i poinformować, jakie techniki mają być stosowane do przygotowywania danych. Może również wpływać na to, które algorytmy mają być stosowane do trenowania modeli uczenia maszynowego.

Image: [https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image-shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU](https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image-shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU)

# Cykl jakości danych – inna perspektywa



**Data profiling**, znane również jako profilowanie danych, to proces analizowania danych w celu zrozumienia ich struktury, zawartości i charakterystyki. Ma to na celu wygenerowanie metadanych, które mogą być wykorzystane do oceny jakości danych, identyfikacji potencjalnych problemów i ułatwienia dostępu do danych.

Istnieje wiele różnych technik data profiling, do których należą:

- Podsumowanie statystyczne
- Analiza typów danych
- Identyfikacja wartości nietypowych
- Analiza rozkładu danych
- Wykrywanie wzorców
- Analiza reguł biznesowych
- Itd..



1.Z. Abedjan, C. Aikora, M. Ouazzani, P. Papotti, and M. Stonebraker. Temporal rules discovery for web data cleaning. Proceedings of the VLDB Endowment (PVLDB), 9(4):336–347, 2015. [\[PDF\]](#) [\[DOI\]](#)

2.Z. Abedjan, L. Goiab, and F. Naumann. Profiling relational data: a survey. VLDB Journal, 24(4):557–581, 2015. [\[PDF\]](#) [\[DOI\]](#)

3.Z. Abedjan, L. Goiab, and F. Naumann. Data profiling (tutorial). In Proceedings of the International Conference on Data Engineering (ICDE), pages 1432–1435, 2016. [\[PDF\]](#) [\[DOI\]](#)

4.Z. Abedjan, T. Grotz, A. Jentsch, and F. Naumann. Profiling and mining RDF data with ProLOD. In Proceedings of the International Conference on Data Engineering (ICDE), pages 1196–1201, 2014. Demo. [\[PDF\]](#) [\[DOI\]](#)

5.Z. Abedjan, J.-A. Quiné-Ruz, and F. Naumann. Detecting unique column combinations on dynamic data. In Proceedings of the International Conference on Data Engineering (ICDE), pages 1036–1047, 2014. [\[PDF\]](#) [\[DOI\]](#)

6.Z. Abedjan, P. Schütze, and F. Naumann. Efficient functional dependency discovery. In Proceedings of the International Conference on Information and Knowledge Management (CIKM), pages 949–958, 2014. [\[PDF\]](#) [\[DOI\]](#)

7.D. Agrawal, P. Berstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Helmy, J. Han, H. V. Jagadish, A. Laney, S. Madden, Y. Papakonstantinou, J. N. Patel, R. Ramakrishnan, K. Ross, C. Shahabi, D. Suciu, S. Venkataraman, and J. Widom. Challenges and opportunities with Big Data. Technical report, Computing Community Consortium, <http://icra.org/coccc/bigdatawhitepaper.pdf>, 2012. [\[PDF\]](#) [\[DOI\]](#)

8.P. Andritsos, R. J. Miller, and P. Tsaparas. Information-theoretic tools for mining database structure from large data sets. In Proceedings of the International Conference on Management of Data (SIGMOD), pages 731–742, 2004. [\[PDF\]](#) [\[DOI\]](#)

9.Apache ATLAS. Data governance and metadata framework for Hadoop. <http://atlas.incubator.apache.org>, 2016. (Online; accessed 25-October-2016). [\[PDF\]](#) [\[DOI\]](#)

10.J. Bauckmann, Z. Abedjan, H. Müller, U. Leser, and F. Naumann. Discovering conditional inclusion dependencies. In Proceedings of the International Conference on Information and Knowledge Management (CIKM), pages 2094–2098, 2012. [\[PDF\]](#) [\[DOI\]](#)

Image: [https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image-shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU](https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image-shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU)

# Cykl jakości danych – inna perspektywa



## Data profiling - Mierniki profilowania danych

Mierniki profilowania danych to wskaźniki, które służą do oceny jakości, charakterystyki i struktury danych.

Istnieje wiele różnych mierników profilowania danych, ale do najczęstszych należą:

- **Statystyki opisowe:** Średnia, Mediana, Odchylenie standardowe, Minimum, Maksimum, Wariancja, Kurtoza, Skośność
- **Analiza typów danych:** Typ danych, Długość pola, Format danych
- **Identyfikacja wartości nietypowych:**
  - Wartości brakujące: Wartości, które nie zostały zarejestrowane.
  - Błędne wartości: Wartości, które są niepoprawne lub nieprawidłowe.
  - Duplikaty: Wartości, które są powtarzane w zbiorze danych.
  - Outliers: Wartości, które znacznie różnią się od innych wartości w zbiorze danych.
- **Analiza rozkładu danych:**
  - Histogramy: Wykresy słupkowe pokazujące częstość występowania różnych wartości w zbiorze danych.
  - Wykresy pudełkowe: Wykresy, które pokazują medianę, kwartyle i zakres wartości w zbiorze danych.
- **Wykrywanie wzorców:**
  - Korelacje: Miary siły i kierunku liniowego związku między dwiema zmiennymi.
  - Zależności: Nieliniowe relacje między zmiennymi.
- **Analiza reguł biznesowych:**
  - Sprawdzanie zgodności: Określanie, czy dane spełniają zdefiniowane reguły biznesowe.
  - Identyfikacja wyjątków: Znajdowanie przypadków, w których dane nie spełniają reguł biznesowych.

Image: [https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image-shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU](https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image-shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU)

# Cykl jakości danych – inna perspektywa



## Data profiling - Mierniki profilowania danych

### Dodatkowe mierniki:

- Liczba rekordów:** Liczba wierszy w tabeli.
- Liczba pól:** Liczba kolumn w tabeli.
- Stosunek pustych wartości:** Procent wartości brakujących w zbiorze danych.
- Stosunek duplikatów:** Procent duplikatów w zbiorze danych.
- Poziom entropii:** Miara uporządkowania danych.

Profilowanie danych jest procesem, który może pomóc w poprawie jakości danych, ułatwić dostęp do danych i udokumentować ich strukturę. Korzystanie z odpowiednich mierników profilowania danych może zapewnić cenne informacje o danych, które mogą być wykorzystane do podejmowania lepszych decyzji biznesowych.

Image: [https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image-shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU](https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image-shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU)

# Cykl jakości danych – inna perspektywa



## Data profiling

Przykładowe narzędzia wykorzystywane do profilowania danych:

- Informatica Data Quality
- IBM InfoSphere Data Quality
- Oracle Data Quality
- SAS Data Quality
- SPSS Data Quality
- R
- Python
- Talend Data Quality, Ataccama Data Quality Center (DQC), Trifacta Wrangler, Alteryx, OpenRefine (dawniej Google Refine), Data Ladder, DataCleaner i wiele innych ...

Image: [https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image-shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU](https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image-shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU)

# Cykl jakości danych – inna perspektywa



## Data profiling - Mierniki profilowania danych

Oprócz wymienionych powyżej mierników, istnieją również bardziej zaawansowane mierniki profilowania danych, które mogą być stosowane w określonych sytuacjach. Na przykład, w przypadku danych tekstowych można stosować techniki przetwarzania języka naturalnego (NLP) do analizy słownictwa, składni i semantyki tekstu, np.:

- Częstotliwość Słów: Analiza ilościowa pokazująca, jak często poszczególne słowa pojawiają się w tekście. Pomaga to zidentyfikować najbardziej dominujące tematy i koncepty.
- Konkordancje: Pokazują, jak słowa są używane w kontekście, co pozwala na zrozumienie znaczenia i użycia w różnych kontekstach.
- Kolokacje: Analiza par lub grup słów, które często występują razem. Może to ujawnić stałe wyrażenia lub specjalistyczną terminologię używaną w określonym obszarze.
- Analiza Sentymentu: Określenie ogólnego nastawienia, emocji lub opinii wyrażanych w tekście. Pozwala na ocenę, czy treść jest pozytywna, negatywna, czy neutralna.
- Ekstrakcja Nazw Własnych: Identyfikacja i kategoryzacja nazw osób, organizacji, miejsc i innych istotnych jednostek.
- Analiza Zależności Składniowych: Analiza struktury zdań, w tym relacji między słowami, aby zrozumieć złożone struktury gramatyczne.
- Topic Modeling: Wykorzystanie algorytmów takich jak Latent Dirichlet Allocation (LDA) do odkrywania ukrytych tematów przewijających się przez zbiory dokumentów tekstowych.
- Term Frequency-Inverse Document Frequency (TF-IDF): Ważenie częstotliwości słów w kontekście ich unikalności w całym korpusie dokumentów, aby znaleźć słowa ważne dla konkretnego dokumentu w stosunku do całej kolekcji.
- Analiza Współwystępowania: Badanie, jak często i w jakich kombinacjach pojawiają się słowa, co może pomóc w mapowaniu związków semantycznych.
- Word Embeddings: Techniki takie jak Word2Vec lub GloVe, które mapują słowa do wektorów liczbowych, reprezentując ich znaczenie i semantyczne relacje w wielowymiarowej przestrzeni.
- Named Entity Recognition (NER): Proces rozpoznawania i klasyfikowania istotnych informacji (nazwane byty) w tekście.
- Syntactic Parsing: Analiza struktury składniowej tekstu, aby zrozumieć hierarchiczne związki między słowami.

$$w_{x,y} = tf_{x,y} \times \log \left( \frac{N}{df_x} \right)$$

**TF-IDF**  
Term  $x$  within document  $y$   
 $tf_{x,y}$  = frequency of  $x$  in  $y$   
 $df_x$  = number of documents containing  $x$   
 $N$  = total number of documents

Image: [https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image\\_shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRjuJUuPjDzWzqBz](https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image_shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRjuJUuPjDzWzqBz)

# Cykl jakości danych – inna perspektywa

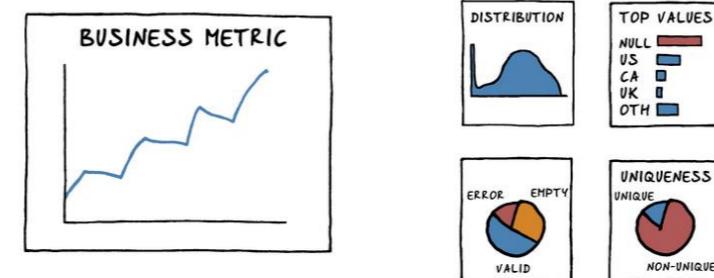


## Data profiling

Istnieją co najmniej dwa typy analizy, które możesz (i powinieneś) zawsze wykonać. Różnią się one *celem* analizy:

- Analiza Danych - analizujesz dane, aby dowiedzieć się czegoś o procesach, które reprezentują, tj. zazwyczaj o twoim biznesie.
- Profilowanie Danych - analizujesz dane, aby dowiedzieć się więcej o samych danych - ich przydatności, jakości, zakresie itp. Profilowanie danych powinno **poprzedzać** analizę danych, abyś mógł upewnić się, że możesz ufać danym.

## DATA ANALYSIS      DATA PROFILING



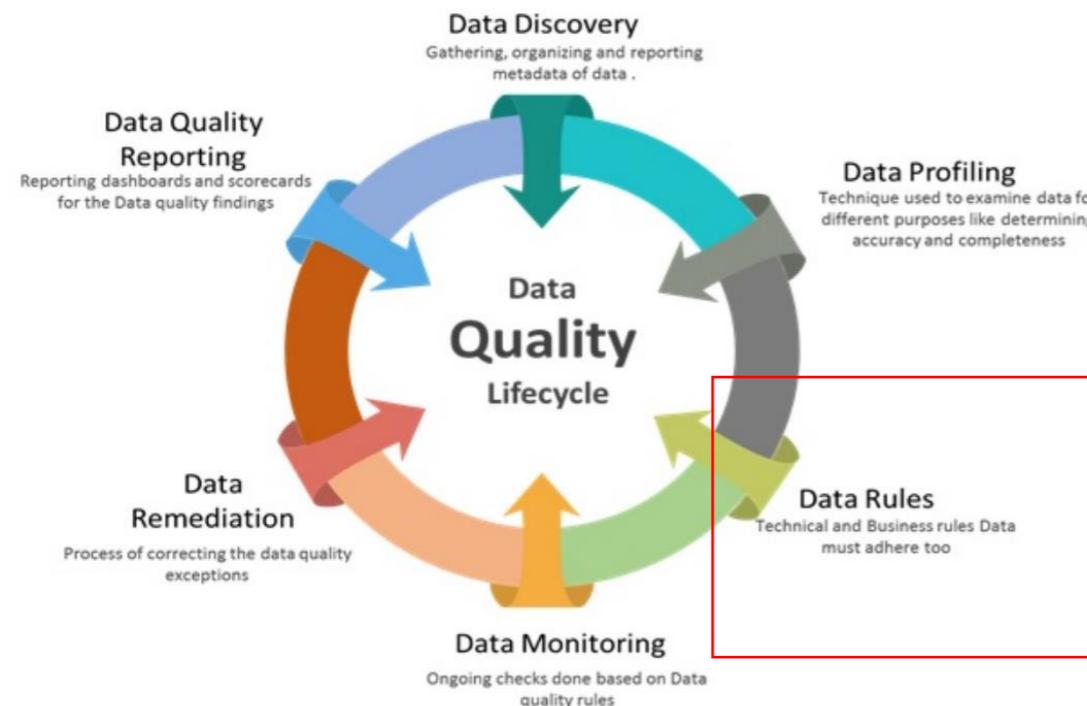
„OUR BUSINESS IS BOOMING!”      „OUR DATA IS C##P...”

 Dataedo /cartoon

Piotr@Dataedo

Image: [https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image-shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU](https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image-shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU)

# Cykl jakości danych – inna perspektywa



Te zasady określają, jak dane są zbierane, przetwarzane, przechowywane i wykorzystywane.

**I. Techniczne zasady** dotyczą specyfikacji i ograniczeń związanych z formatem, strukturą i procesowaniem danych (np. Formatowanie i Walidacja, Normalizacja, Integracja, Szyfrowanie i Bezpieczeństwo, Backup i Odtwarzanie)

**II. Biznesowe zasady** odnoszą się do wymagań i celów organizacji, które wpływają na zarządzanie danymi:

- dostępność: Dane muszą być dostępne dla uprawnionych użytkowników, kiedy są potrzebne, bez nieuzasadnionych opóźnień
- spójność: Dane w całej organizacji powinny być spójne i odzwierciedlać rzeczywisty stan rzeczy.
- aktualność: Dane muszą być aktualne i odświeżane w regularnych odstępach czasu, aby odzwierciedlać najnowszy stan.
- prywatność i zgodność: Przetwarzanie danych musi być zgodne z obowiązującymi przepisami dotyczącymi prywatności i ochrony danych, takimi jak np. GDPR (*ang. General Data Protection Regulation*) w Europie.
- jakość i dokładność: Muszą być wdrożone procesy zapewnienia jakości, aby dane były dokładne, kompleksowe i wolne od błędów.
- zarządzanie Życiem Cyklu Danych: Definiowanie, jak długo dane są przechowywane, kiedy są archiwizowane lub usuwane, zgodnie z wymaganiami biznesowymi i prawnymi.
- zarządzanie Metadanymi: Metadane opisujące dane muszą być utrzymywane, aby ułatwić ich katalogowanie, wyszukiwanie i zarządzanie.

Image: [https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image-shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU](https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image-shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU)

# Cykl jakości danych – inna perspektywa



Monitorowanie danych, czyli ciągłe kontrolowanie jakości danych. Zapewnia ono, że dane pozostają dokładne, spójne i użyteczne w czasie. Proces ten obejmuje systematyczne sprawdzanie danych względem wcześniej zdefiniowanych wskaźników.

## Składniki Monitorowania Danych

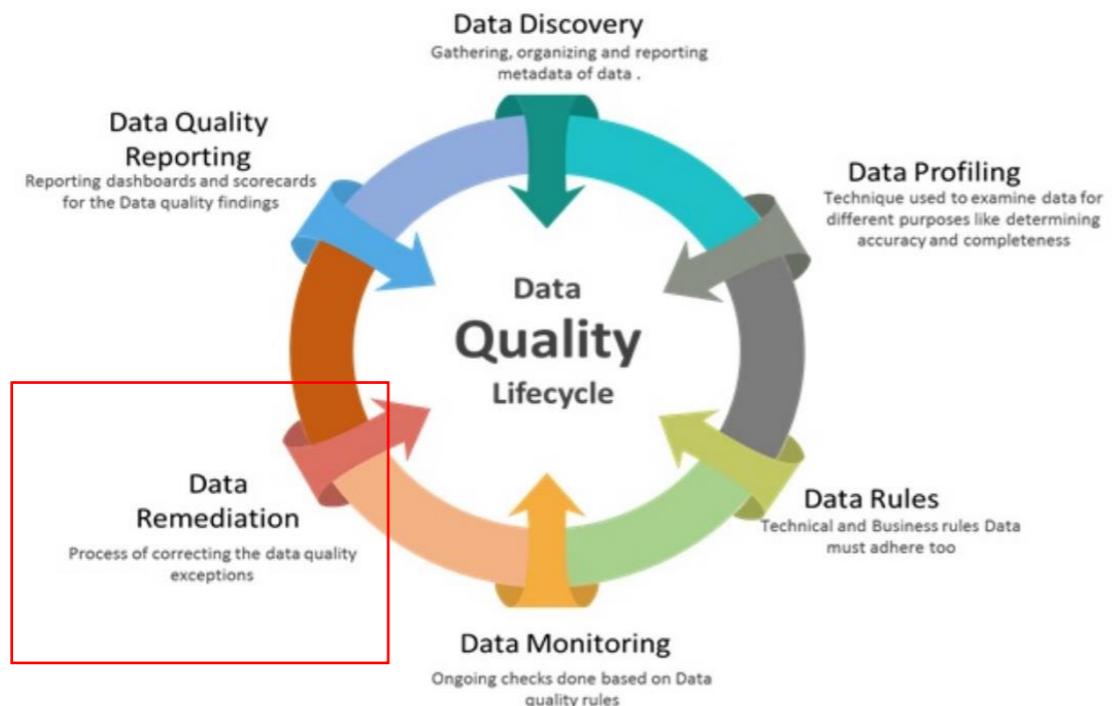
- Zasady i Metryki Jakości:** Definiowanie konkretnych, mierzalnych kryteriów, którym dane muszą odpowiadać, takich jak dokładność, kompletność, spójność, aktualność i unikalność.
- Narzędzia do Automatycznego Monitorowania:** Wykorzystanie narzędzi do automatycznego skanowania, analizy i raportowania jakości danych, znacznie zwiększające efektywność i zakres działań monitorujących.
- Dashboardy i Alerty:** Implementacja dashboardów w czasie rzeczywistym i mechanizmów powiadamiania, aby zapewnić widoczność jakości danych i informować odpowiednie strony o problemach w miarę ich pojawiania się.
- Regularne Audyty:** Planowanie okresowych audytów manualnych, które uzupełniają monitorowanie automatyczne, pozwalając na głębsze zanurzenie w problemy z jakością danych i skuteczność trwających praktyk.

Przykładowe narzędzia dla różnych zastosowań biznesowych:

- Microsoft Power BI - Narzędzie do wizualizacji danych, które również oferuje monitoring i analizę danych w czasie rzeczywistym (np. Analizator Wydajności)
- Google Analytics - Powszechnie używany do śledzenia i analizowania ruchu na stronach internetowych, zachowania użytkowników i konwersji.
- Prometheus - System monitorowania i alertowania otwartego kodu, który jest często używany w środowiskach DevOps.
- Zabbix - Oprogramowanie open-source do monitorowania sieci, serwerów, wirtualnych maszyn i innych urządzeń sieciowych.
- Nagios - System monitorowania, który może wykrywać i rozwiązywać problemy w infrastrukturze IT przed wpływem na krytyczne procesy biznesowe.
- Splunk - Oferuje funkcjonalność do monitorowania, przeszukiwania, analizy i wizualizacji danych maszynowych i logów w czasie rzeczywistym.
- Datadog - Platforma monitorowania, która zapewnia wgląd w wydajność aplikacji, infrastruktury i usług w chmurze.
- ELK Stack (Elasticsearch, Logstash, Kibana) - Służy do monitorowania logów, analizy danych i ich wizualizacji.

Image: [https://media.licdn.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image-shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU](https://media.licdn.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image-shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU)

# Cykl jakości danych – inna perspektywa

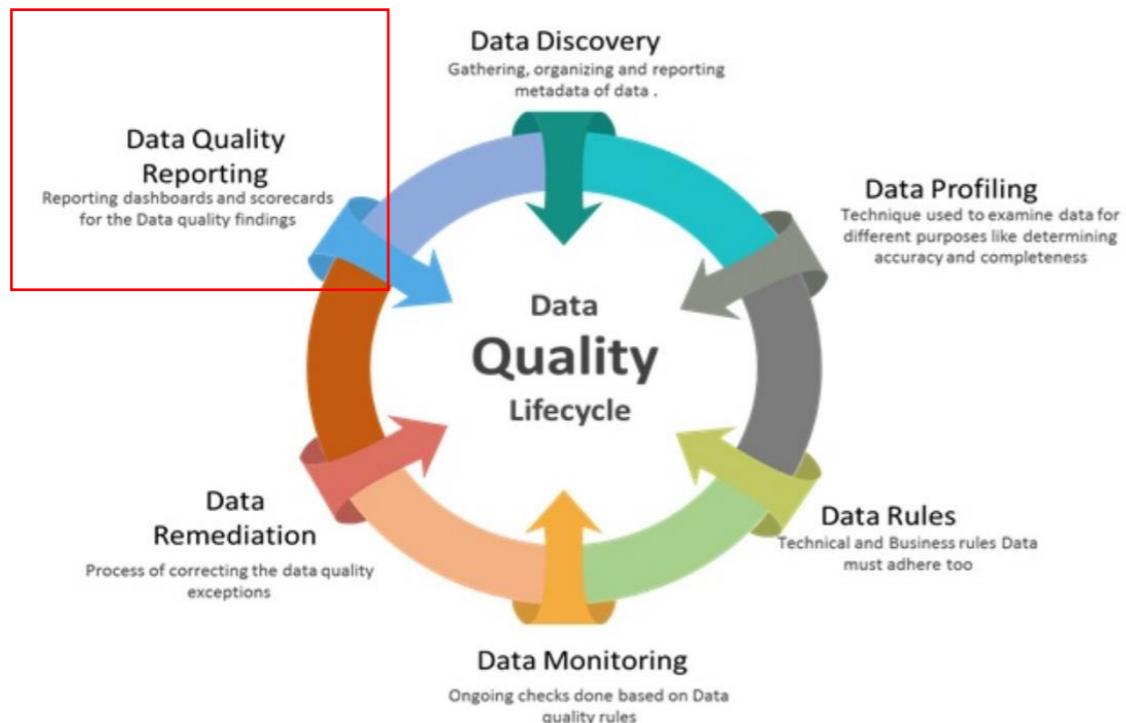


Data Remediation (inaczej korygowanie danych, remediacja danych, dot, data governance) - proces podejmowany w celu naprawienia problemów z jakością danych, które zostały zidentyfikowane podczas fazy oceny jakości danych. Jest to kluczowy element ram zarządzania danymi, aby zapewnić, że dane w organizacji są dokładne, spójne i wiarygodne. Proces remediacji zazwyczaj obejmuje kilka kluczowych kroków:

- **Identyfikacja:** Polega na wykrywaniu problemów z jakością danych poprzez profilowanie danych, audyty lub systemy monitorowania (kroki wcześniejsze)
- **Ocena:** Po zidentyfikowaniu problemów z jakością danych, każdy problem jest oceniany pod kątem jego wpływu na operacje biznesowe i podejmowanie decyzji. Ten krok pomaga ustalić priorytety działań remediacji na podstawie wagi i pilności.
- **Analiza Przyczynowa:** Zrozumienie, dlaczego doszło do problemów z jakością danych, jest niezbędne. Mogło to być spowodowane błędami wprowadzania danych, nieprawidłowym mapowaniem danych, problemami z integracją systemów lub brakiem standardowych procedur operacyjnych.
- **Korekta:** To właściwy krok remediacji, gdzie korygowane są zidentyfikowane problemy. Korekty mogą być ręczne, takie jak ponowne wprowadzenie danych, lub automatyczne, jak uruchomienie skryptów w celu naprawienia rozpoznanych i jednolitych błędów.
- **Zapobieganie:** Po skorygowaniu problemów ważne jest, aby zająć się przyczynami podstawowymi, aby zapobiec ponownemu wystąpieniu tych samych problemów w przyszłości. Może to obejmować zmianę procesów, aktualizację protokołów wprowadzania danych, poprawę reguł validacji danych lub lepsze szkolenie personelu.
- **Dokumentacja:** Dokumentowanie procesu remediacji, w tym informacji o tym, co zostało poprawione, jak i dlaczego, zapewnia przejrzystość i pomaga w udoskonalaniu ciągłej strategii jakości danych.
- **Monitorowanie Po Remediacji:** Po remediacji istotne jest kontynuowanie monitorowania danych, aby upewnić się, że korekty są trwałe i że nie wprowadzono nowych problemów.
- **Raportowanie:** Komunikowanie wyników działań remediacji interesariuszom jest niezbędne. Obejmuje to szczegółowe opisanie podjętych działań, rezultatów oraz wszelkich zaleceń dotyczących dalszych ulepszeń.

Przykładowe narzędzia: **Talend Data Quality**, **Ataccama Data Quality Center (DQC)**, **Trifacta Wrangler**, **Alteryx**, **OpenRefine (dawniej Google Refine)**, **Data Ladder**, **DataCleaner** i wiele innych ...

# Cykl jakości danych – inna perspektywa



**Raportowanie jakości danych** – (czyli raportujemy nie tylko procesy jakie dane opisują, ale także możemy samą ich wartość) - proces prezentacji wyników z działań związanych z jakością danych, który może obejmować zarówno statyczne raporty, jak i dynamiczne dashboardy oraz karty wyników (scorecards). Jest to kluczowy element zarządzania jakością danych, ponieważ umożliwia interesariuszom zrozumienie obecnego stanu danych i podejmowanie świadomych decyzji w oparciu o te informacje.

Dashboard (panel kontrolny) jakości danych to interaktywna aplikacja, która prezentuje w czasie rzeczywistym kluczowe wskaźniki jakości danych (Key Quality Indicators - KQIs) i pozwala użytkownikom na głębszą analizę danych poprzez wiercenie (drill down) w prezentowanych danych.

Dashboardy mogą zawierać:

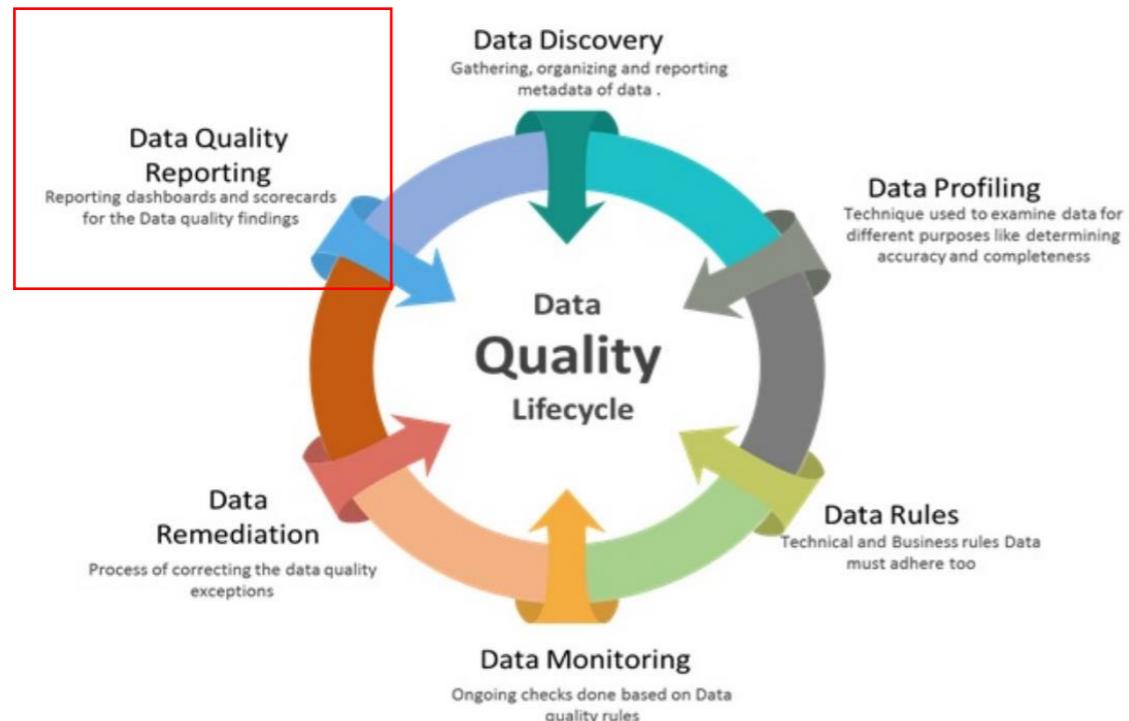
1. Wskaźniki Jakości: Wizualne przedstawienie metryk jakości, takich jak dokładność, kompletność, spójność, aktualność i unikalność danych.
2. Trendy w Czasie: Wykresy pokazujące, jak wskaźniki jakości danych zmieniają się w czasie.
3. Alarmy i Powiadomienia: Interaktywne alerty wskazujące na obszary, które wymagają uwagi lub przekroczyły ustalone progi.
4. Szczegółowe Raporty: Możliwość przejścia do szczegółowych raportów dotyczących konkretnych problemów z danymi.



Dużo *templatów* jako inspiracje w Internecie ...

Image: [https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image-shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU](https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image-shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU)

# Cykl jakości danych – inna perspektywa



## Karty Wyników Jakości Danych (Data Quality Scorecards)

Karty wyników to zbiory metryk, które są raportowane w regularnych odstępach czasu i dostarczają podsumowanie ogólnej wydajności danych względem z góry ustalonych celów jakościowych. Karty wyników mogą zawierać:

1. Oceny Jakości: Numeryczne lub procentowe oceny określające poziom spełnienia wymagań jakościowych dla różnych aspektów danych.
2. Porównanie z Celami: Wizualizacja porównująca obecne wyniki z założonymi celami lub standardami branżowymi.
3. Postępy w Czasie: Prezentacja poprawy lub pogorszenia jakości danych w kolejnych okresach.
4. Przegląd Obszarów Danych: Segmentacja danych w różne obszary, takie jak klient, produkt, transakcja itp., z oceną jakości dla każdego segmentu.

Dużo *templat'ów* jako inspiracje w Internecie ...

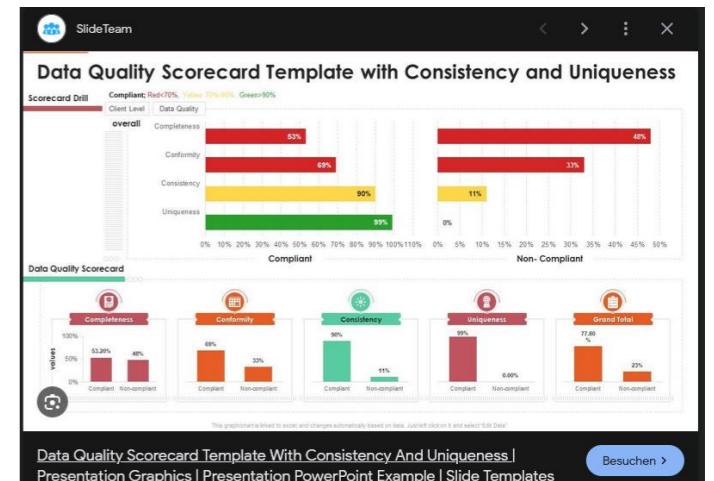
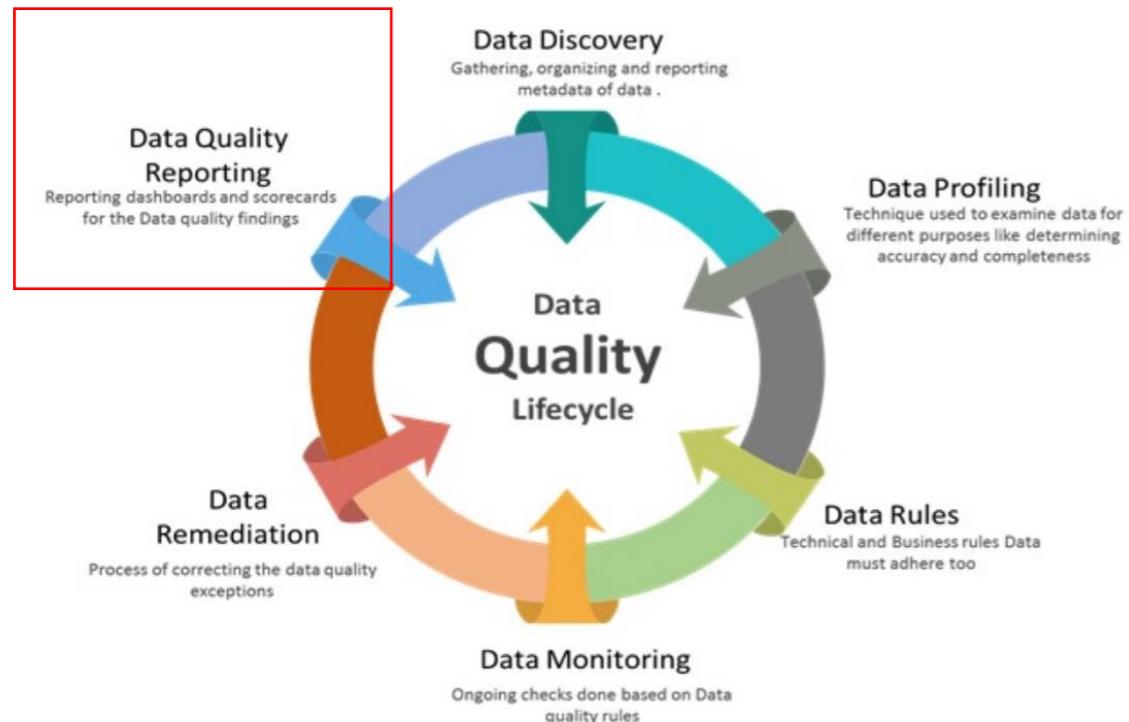


Image: [https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image-shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU](https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image-shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU)

# Cykl jakości danych – inna perspektywa



## Przygotowanie Raportów i Dashboardów

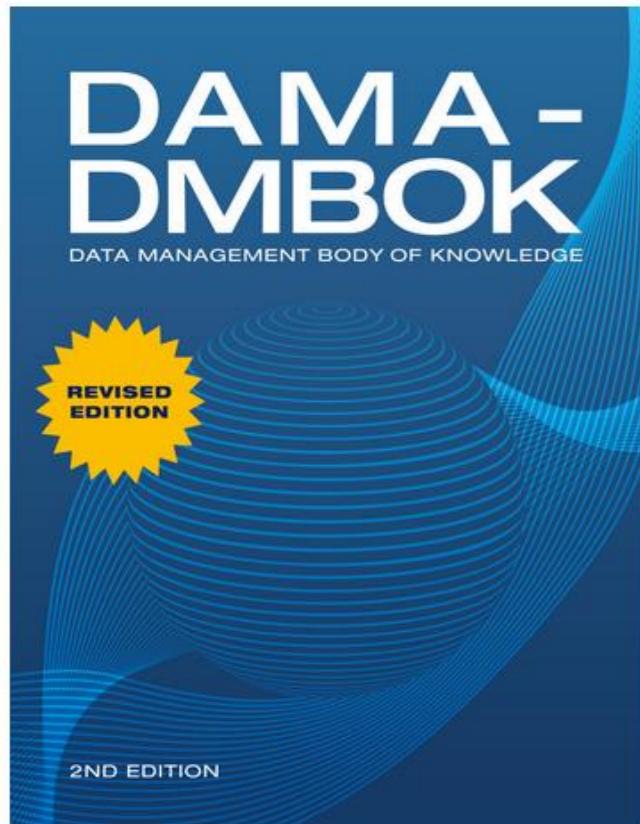
Przy tworzeniu raportów i dashboardów jakości danych, należy wziąć pod uwagę:

1. Odbiorcy: Zrozumienie, kto będzie korzystać z raportów/dashboardów, i dostosowanie prezentacji do ich potrzeb.
2. Cel: Jasne określenie, co raport lub dashboard ma przekazać i jakie decyzje ma wspierać.
3. Dane: Upewnienie się, że dane podstawowe są dokładne i aktualne.
4. Narzędzia: Wybór narzędzi do raportowania i wizualizacji danych, które najlepiej spełniają wymagania organizacji, np. Tableau, Power BI, Qlik, itd.

Skuteczne raportowanie jakości danych umożliwia organizacjom śledzenie wpływu danych na cele biznesowe, identyfikowanie obszarów wymagających uwagi, a także świadczy o dojrzałości organizacji w zakresie zarządzania danymi.

Image: [https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover\\_image-shrink\\_720\\_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU](https://media.linkedin.com/dms/image/C5112AQEA8MyLMsQX8g/article-cover_image-shrink_720_1280/0/1567931089544?e=2147483647&v=beta&t=CV-4MXH5sRIJuDTGQ9fSqnHYApYLWbXrsQnysAinonU)

# Cykl danych - DATA MANAGEMENT BODY OF KNOWLEDGE



DATA MANAGEMENT BODY OF KNOWLEDGE

## DAMA-DMBOK2

DAMA-DMBOK2 was originally published in 2018 in English and has since been translated into multiple languages.

DAMA International is pleased to bring you the Revised Edition, released in March 2024.  
[DAMA-DMBOK2 Revised Edition FAQs](#)  
[Significant Changes to DAMA-DMBOK2](#)

The contributors are all experienced practitioners with names you may recognize. This is not a theoretical book, although it has authoritative theoretical substance. It is primarily a book of practice, experience, and expression of what works by the very best practitioners in the industry today.

LEARN MORE

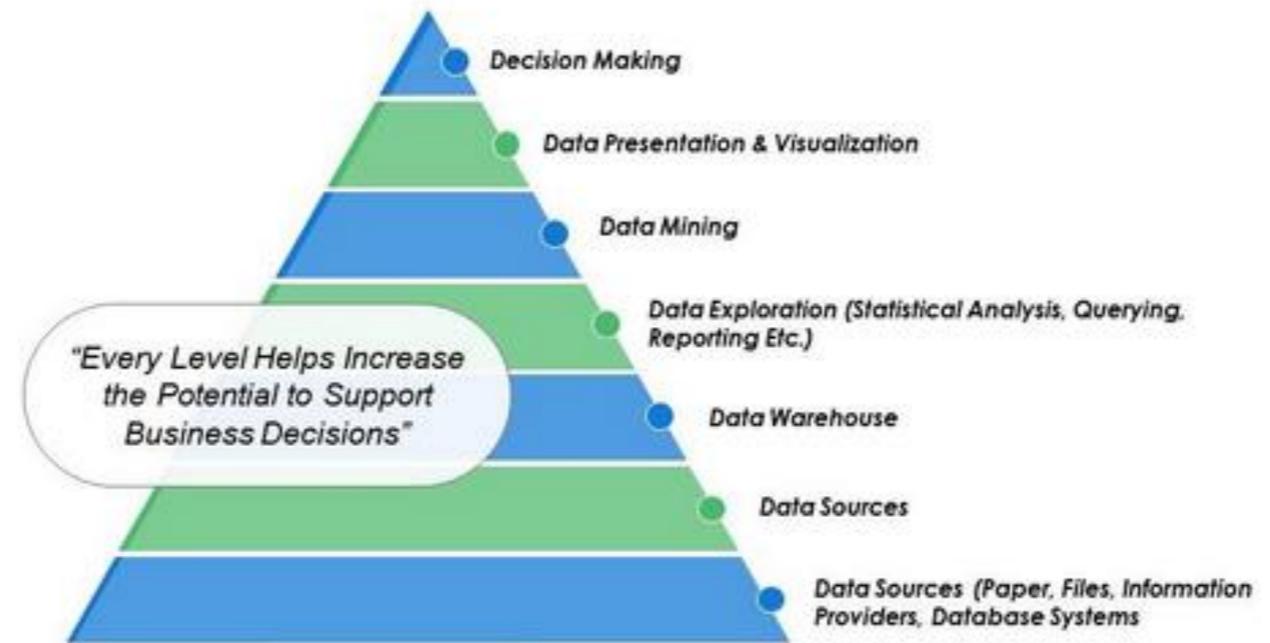
[www.dama.org](http://www.dama.org)

[agh.edu.pl](http://agh.edu.pl)

DAMA-DMBOK, czyli Data Management Body of Knowledge opracowany przez Data Management Association International (DAMA International), to ramy i podręcznik dobrej praktyki dla zarządzania danymi w organizacjach. Założenia DAMA-DMBOK dotyczą tematów takich jak:

- Zarządzanie danymi jako dyscyplina:** Traktowanie zarządzania danymi jako zintegrowanej dyscypliny zarządzania, a nie jako zbioru niezależnych praktyk.
- Funkcje zarządzania danymi:** Wyróżnienie kluczowych funkcji zarządzania danymi, takich jak zarządzanie jakością danych, zarządzanie metadanymi, zarządzanie danymi referencyjnymi i master, zarządzanie danymi osobowymi, zarządzanie architekturą danych, zarządzanie bezpieczeństwem danych, itd.
- Zarządzanie jako klucz do wartości biznesowej:** Podkreślenie, że efektywne zarządzanie danymi jest niezbędne do wykorzystania danych jako aktywa, które generują wartość biznesową.
- Proces ciągłego doskonalenia:** Rozpoznanie, że zarządzanie danymi to proces ciągłego doskonalenia, który powinien być dostosowywany w miarę zmieniających się potrzeb biznesowych i technologii.
- Standardy i najlepsze praktyki:** Zastosowanie standardów i najlepszych praktyk w zakresie zarządzania danymi w celu osiągnięcia wysokiej jakości, efektywności operacyjnej i zgodności regulacyjnej.
- Udział interesariuszy:** Angażowanie interesariuszy z różnych poziomów organizacji, w tym kierownictwa, użytkowników biznesowych i IT, w celu wspierania celów zarządzania danymi.
- Edukacja i rozwój:** Promowanie edukacji i rozwoju w dziedzinie zarządzania danymi, aby budować wiedzę i kompetencje wewnętrz organizacji.
- Zarządzanie zmianą:** Zarządzanie zmianą jako kluczowy element wprowadzania i utrzymywania skutecznych praktyk zarządzania danymi.
- Ocena i pomiar:** Regularne ocenianie i mierzenie efektywności praktyk zarządzania danymi w celu ich optymalizacji.
- Zintegrowane podejście:** Postrzeganie zarządzania danymi jako części większego ekosystemu organizacji, zrozumienie i wspieranie wzajemnych zależności między danymi a procesami biznesowymi.
11. I wiele wiele innych – warto!

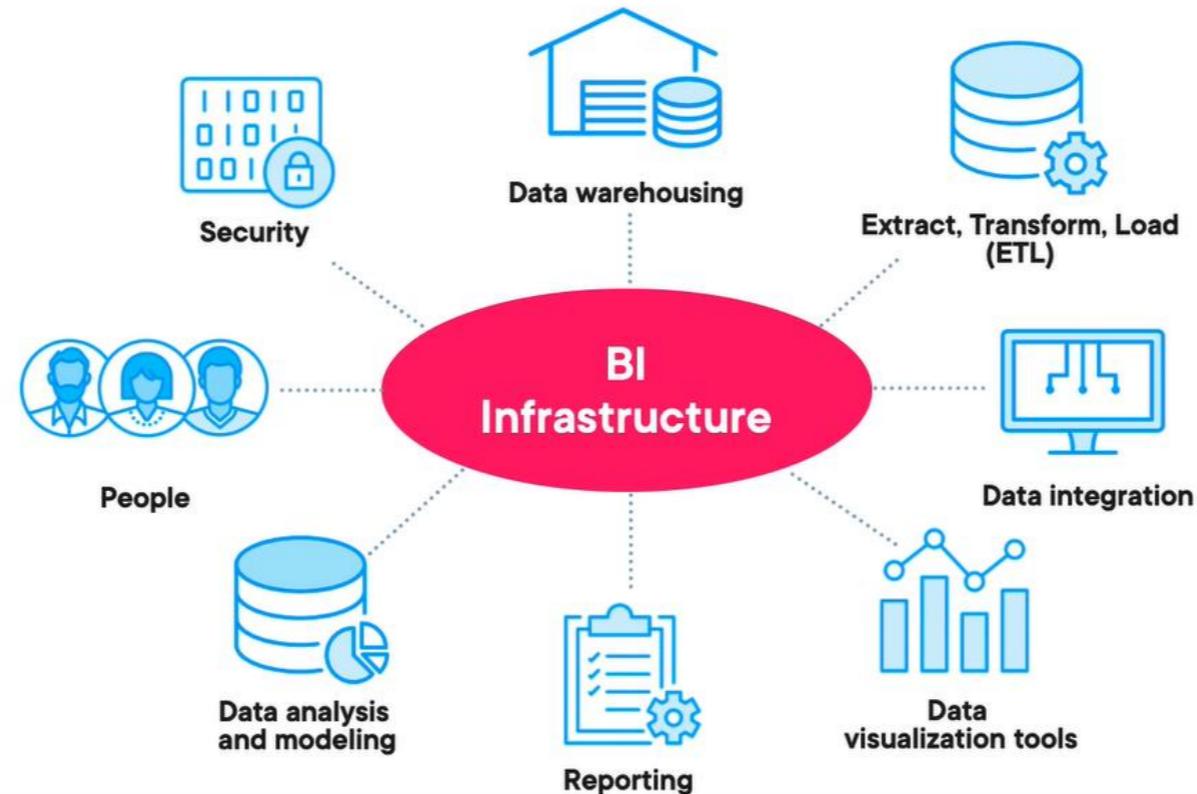
Poruszyliśmy wiele perspektyw, wiele tematów ...



Graphic source: <https://www.slideteam.net/business-intelligence-process-pyramid.html>

... powrót do podstawowych koncepcji

Poruszyliśmy wiele perspektyw, wiele tematów ...



Graphic: Jamie Champagne

... powrót do podstawowych koncepcji

Zaczniemy od pierwszego i często największego pytania: gdzie zamierzasz umieścić te wszystkie dane?  
Odpowiedź brzmi – to zależy, np. w hurtowni danych.

Struktura zarządzania danymi będzie często zależała od ich rozmiaru, ale zwykle wykorzystuje się komponenty takie jak bazy danych, magazyny danych, a nawet tzw. jeziora danych, aby gromadzić te informacje.

Należy dokładnie rozważyć swoje źródła i sposób pozyskiwania danych do swojego centralnego repozytorium danych biznesowych.

Gdy dane mają być skoncentrowane lub przynajmniej pobrane do jednego źródła, wprowadzany jest proces ETL ekstrakcji, transformacji i ładowania danych.



## Extract, Transform, Load (ETL)

Zacznijmy od pierwszego i często największego pytania: gdzie zamierzasz umieścić te wszystkie dane?  
Odpowiedź brzmi – to zależy, np. w hurtowni danych.

## Extract, Transform, Load (ETL)

Proces ETL to skrót od "Ekstrakcja, Transformacja i Ładowanie" (ang. Extract, Transform, Load). Jest to proces wykorzystywany w dziedzinie informatyki i analizy danych, którego celem jest przeniesienie danych z różnych źródeł do docelowego magazynu danych lub bazy danych w sposób skuteczny i spójny.



### **Ekstrakcja (Extract):**

Polega na pobieraniu danych z różnych źródeł, takich jak bazy danych, pliki CSV, Excel, API, czy struktury webowe.

### **Transformacja (Transform):**

W tym etapie dane są poddawane różnym procesom transformacji, takim jak oczyszczanie danych, usuwanie duplikatów, normalizacja danych, obliczenia pochodne, standaryzacja formatów itp.

### **Ładowanie (Load):**

W tym etapie przetworzone i przygotowane dane są ładowane do docelowego magazynu danych, gdzie mogą być wykorzystane do analizy biznesowej, raportowania i generowania wniosków.

Zaczniemy od pierwszego i często największego pytania: gdzie zamierzasz umieścić te wszystkie dane?  
Odpowiedź brzmi – to zależy, np. w hurtowni danych.

## Extract, Transform, Load (ETL)

Istnieje wiele narzędzi do przeprowadzania procesu ETL (Ekstrakcja, Transformacja i Ładowanie) danych. Oto kilka popularnych narzędzi ETL:



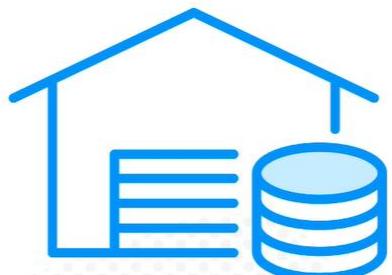
- **Apache Nifi:** Otwarte źródło, łatwe w użyciu narzędzie do automatyzacji przepływu danych, umożliwiające przetwarzanie strumieniowe i wsadowe danych.
- **Talend Open Studio:** Bezpłatne narzędzie ETL z obsługą graficznego interfejsu użytkownika, umożliwiające łatwe projektowanie, testowanie i wdrażanie procesów ETL.
- **Informatica PowerCenter:** Zintegrowane narzędzie ETL, które oferuje zaawansowane funkcje do zarządzania danymi, integrowania różnych źródeł danych i tworzenia zaawansowanych transformacji.
- **Microsoft SQL Server Integration Services (SSIS):** Narzędzie ETL dostępne w ramach pakietu Microsoft SQL Server, umożliwiające projektowanie, wdrażanie i zarządzanie procesami ETL w środowisku Windows.
- **IBM InfoSphere DataStage:** Potężne narzędzie ETL opracowane przez IBM, które zapewnia wsparcie dla przetwarzania dużych ilości danych, integracji wielu źródeł danych i tworzenia zaawansowanych transformacji.
- **Pentaho Data Integration (Kettle):** narzędzie ETL oferujące bogaty zestaw funkcji do ekstrakcji, transformacji i ładowania danych, dostępne jako część platformy Business Intelligence Pentaho.
- **SAP Data Services:** Zaawansowane narzędzie ETL opracowane przez SAP, które umożliwia integrację danych, przekształcanie strumieniowe i wsadowe, oraz zarządzanie jakością danych.
- **Matillion ETL:** Narzędzie ETL zbudowane na platformie chmurowej, które oferuje szybkie i skalowalne przetwarzanie danych w środowisku chmurowym takim jak Amazon Web Services (AWS) lub Google Cloud Platform (GCP).
- **SAS Data Management:** Kompleksowe narzędzie do zarządzania danymi, które obejmuje funkcje ETL, jakości danych, integracji danych i zarządzania metadanymi.
- **Apache Spark:** Otwarte źródło platformy przetwarzania danych, która oferuje moduł Spark SQL, który można wykorzystać do przeprowadzania operacji ETL na dużych zbiorach danych w czasie rzeczywistym.
- I wiele innych ...

Narzędzia ETL różnią się pod względem funkcji, elastyczności, złożoności i kosztów, więc warto zrozumieć wymagania swojego projektu, aby wybrać narzędzie najlepiej dopasowane do potrzeb.

Zaczniemy od pierwszego i często największego pytania: gdzie zamierzasz umieścić te wszystkie dane?  
Odpowiedź brzmi – to zależy, np. w hurtowni danych.

## Extract, Transform, Load (ETL)

Istnieje wiele narzędzi do przeprowadzania procesu ETL (Ekstrakcja, Transformacja i Ładowanie) danych. Oto kilka popularnych narzędzi ETL:



- **Apache Nifi:** Otwarte źródło, łatwe w użyciu narzędzie do automatyzacji przepływu danych, umożliwiające przetwarzanie strumieniowe i wsadowe danych.
- **Talend Open Studio:** Bezpłatne narzędzie ETL z obsługą graficznego interfejsu użytkownika, umożliwiające łatwe projektowanie, testowanie i wdrażanie procesów ETL.
- **Informatica PowerCenter:** Zintegrowane narzędzie ETL, które oferuje zaawansowane funkcje do zarządzania danymi, integrowania różnych źródeł danych i tworzenia zaawansowanych transformacji.
- **Microsoft SQL Server Integration Services (SSIS):** Narzędzie ETL dostępne w ramach pakietu Microsoft SQL Server, umożliwiające projektowanie, wdrażanie i zarządzanie procesami ETL w środowisku Windows.
- **IBM InfoSphere DataStage:** Potężne narzędzie ETL opracowane przez IBM, które zapewnia wsparcie dla przetwarzania dużych ilości danych, integracji wielu źródeł danych i tworzenia zaawansowanych transformacji.
- **Pentaho Data Integration (Kettle):** narzędzie ETL oferujące bogaty zestaw funkcji do ekstrakcji, transformacji i ładowania danych, dostępne jako część platformy Business Intelligence Pentaho.
- **SAP Data Services:** Zaawansowane narzędzie ETL opracowane przez SAP, które umożliwia integrację danych, przekształcanie strumieniowe i wsadowe, oraz zarządzanie jakością danych.
- **Matillion ETL:** Narzędzie ETL zbudowane na platformie chmurowej, które oferuje szybkie i skalowalne przetwarzanie danych w środowisku chmurowym takim jak Amazon Web Services (AWS) lub Google Cloud Platform (GCP).
- **SAS Data Management:** Kompleksowe narzędzie do zarządzania danymi, które obejmuje funkcje ETL, jakości danych, integracji danych i zarządzania metadanymi.
- **Apache Spark:** Otwarte źródło platformy przetwarzania danych, która oferuje moduł Spark SQL, który można wykorzystać do przeprowadzania operacji ETL na dużych zbiorach danych w czasie rzeczywistym.
- I wiele innych ...

Narzędzia ETL różnią się pod względem funkcji, elastyczności, złożoności i kosztów, więc warto zrozumieć wymagania swojego projektu, aby wybrać narzędzie najlepiej dopasowane do potrzeb.

Czyli jakie te dane ?

*Non-structured, Semi-structured and Structured data*

Niestrukturyzowane, częściowo ustrukturyzowane, ustrukturyzowane

Czyli jakie te dane ?

Dane to zbiór faktów, takich jak liczby, słowa, pomiary lub po prostu opisy rzeczy.

Id_kota	Imię	Wiek	Płeć	Rasa	Kolor	Waga
1	Mruczek	3	M	Maine Coon	Szary	7.5 kg
2	Luna	2	Ż	Sfinks	Bezsiwy	4.2 kg
3	Simba	5	K	Pers	Pomarańczowy	5.8 kg
4	Pucio	1	M	Scottish Fold	Biały	3.5 kg

## Ustrukturyzowane

Czyli jakie te dane ?

Id_kota	Imię	Wiek	Płeć	Rasa	Kolor	Waga
1	Mruczek	3	M	Maine Coon	Szary	7.5
2	Luna	2	Ż	Sfinks	Bezsiwy	4.2
3	Simba	5	K	Pers	Pomarańczowy	5.8
4	Pucio	1	M	Scottish Fold	Biały	3.5

- **Id\_kota:** jest to atrybut **ilościowy**, ponieważ jest to unikalny identyfikator kota, który służy do jednoznacznego identyfikowania każdego rekordu w bazie danych. W schemacie relacyjnej bazy danych byłoby to klucz główny.
- **Imię:** jest to atrybut **jakościowy**, ponieważ opisuje cechę kota, która nie jest wyrażona w liczbach, ale w postaci ciągu znaków. Jest to również atrybut kandydujący na klucz główny, ale nie jest unikalny dla każdego rekordu, ponieważ dwa koty mogą mieć takie samo imię.
- **Wiek:** jest to atrybut **ilościowy**, ponieważ opisuje liczbę lat kotów. W tym przypadku liczba jest wyrażona w formie liczby całkowitej.
- **Płeć:** jest to atrybut **jakościowy**, ponieważ opisuje cechę kota, która może przyjąć tylko jedną z kilku możliwych wartości (M - samiec, K - samica).
- **Rasa:** jest to atrybut **jakościowy**, ponieważ opisuje cechę kota, która jest wyrażona w postaci ciągu znaków.
- **Kolor:** jest to atrybut **jakościowy**, ponieważ opisuje cechę kota, która jest wyrażona w postaci ciągu znaków.
- **Waga:** jest to atrybut **ilościowy**, ponieważ opisuje wagę kota, wyrazoną w formie liczby zmiennoprzecinkowej.



Qualitative

Quantitative

Niestrukturyzowane

Czyli jakie te dane ?

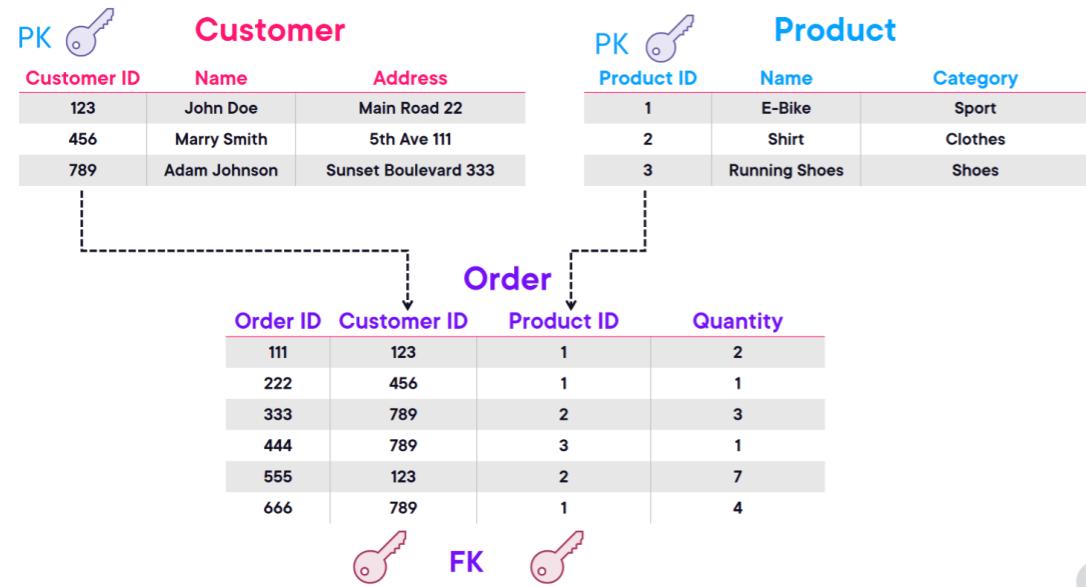
Id_kota	Imię	Wiek	Płeć	Kolor	Waga	<i>Portret</i> 
1	Mruczek	3	M	Szary	7.5 kg	
2	Luna	2	Ż	Czarny	4.2 kg	
3	Simba	5	K	Rudy	5.8 kg	
4	Pucio	1	M	Biały	3.5 kg	

## *Non-structured, Semi-structured and Structured data*

### Niestrukturyzowane, częściowo ustrukturyzowane, ustrukturyzowane

- Tekst wolny: Jest to najbardziej powszechny rodzaj danych niestrukturyzowanych, który może zawierać różnorodne informacje w postaci swobodnie pisanych artykułów, e-maili, tweetów, dokumentów tekstowych itp.
  - Obrazy i multimedia: Dane w postaci obrazów, plików dźwiękowych i wideo są zazwyczaj niestrukturyzowane, chociaż mogą zawierać metadane lub informacje opisowe.
  - Pliki binarne: Dane w formie binarnej, takie jak pliki wykonywalne, pliki graficzne, pliki audio i wideo, nie posiadają spójnej struktury czy znaczników, które można by łatwo zinterpretować.
  - Dane geoprzestrzenne: Dane dotyczące lokalizacji, map, współrzędnych geograficznych itp. mogą być niestrukturyzowane, jeśli nie mają spójnej struktury lub formatu.
  - Pliki PDF: Chociaż pliki PDF mogą zawierać pewne stopnie struktury, często zawierają one różnorodne dane w postaci tekstu, obrazów i innych elementów, które są trudne do analizy automatycznej.
  - Dane z mediów społecznościowych: Posty, komentarze, multimedia i inne treści z platform społecznościowych mogą być niestrukturyzowane, ponieważ zawierają różnorodne informacje w różnych formatach i stylach.
  - Itp..
- 
- **Format JSON (JavaScript Object Notation):** Dane w formacie JSON są elastyczne i mogą zawierać różne rodzaje informacji w formie obiektów i tablic.
  - **Format XML (eXtensible Markup Language):** Dane XML są hierarchicznie strukturyzowane za pomocą znaczników, ale mogą mieć różną głębokość i złożoność struktury.
  - **Dane w formacie CSV (Comma-Separated Values):** Chociaż dane CSV mogą być strukturyzowane w postaci tabelarycznej, mogą również zawierać niestrukturyzowane pola tekstowe.
  - **Dokumenty HTML:** Strony internetowe zawierające dane mogą być również traktowane jako dane półstrukturyzowane, ponieważ mogą zawierać różnorodne tagi i znaczniki, które nadają strukturę, ale mogą też zawierać niestrukturyzowane treści.
  - **Dokumenty tekstowe z wbudowanymi tagami lub metadanymi:** Dokumenty tekstowe mogą zawierać pewne stopnie struktury, na przykład w postaci sekcji, tytułów, list itp., ale również pozwalają na wprowadzenie niestrukturyzowanych danych.
  - **Protokoły logowania (np. logi serwerów):** Dane z logów serwerów mogą być półstrukturyzowane, ponieważ zawierają informacje o zdarzeniach w określonym formacie, ale mogą też zawierać niestrukturyzowane komunikaty lub dodatkowe metadane.
  - **Itp..**
- 
- **Tabele relacyjne:** Dane przechowywane są w postaci tabeli, z wierszami i kolumnami, które są logicznie powiązane ze sobą za pomocą kluczy.
  - **Widoki:** Widoki są logicznymi zestawami danych pochodzącymi z jednej lub więcej tabel. Są one używane do uproszczenia złożonych zapytań lub prezentowania danych w bardziej przystępny sposób.
  - **Indeksy:** Indeksy są strukturami danych, które przyspieszają dostęp do danych, identyfikując klucze i przyspieszając wyszukiwanie.
  - **Zarządzane procedury składowane (Stored Procedures):** Procedury składowane to zbiorowe instrukcje SQL przechowywane i wykonane na serwerze bazy danych.
  - **Widoki materiałowane (Materialized Views):** Widoki materiałowane są fizycznie przechowywanymi wynikami zapytań, co pozwala na szybszy dostęp do danych, ale wymaga aktualizacji w przypadku zmian danych źródłowych.
  - **Schematy i metadane:** Schematy określają strukturę danych w magazynie danych, a metadane zawierają informacje opisujące dane, takie jak nazwy kolumn, typy danych, klucze itp. Kolumny obliczane (Computed Columns): Kolumny obliczane są generowane na podstawie innych kolumn w tabeli, na przykład poprzez obliczenia matematyczne lub konkatenację ciągów znaków.
  - **Klucze główne i obce:** Klucze główne są unikatowymi identyfikatorami dla każdego wiersza w tabeli, podczas gdy klucze obce określają powiązania między tabelami.
  - **Itp..**

# Dane ustrukturyzowane



Wszystkie wiersze tabeli mają ten sam zestaw kolumn

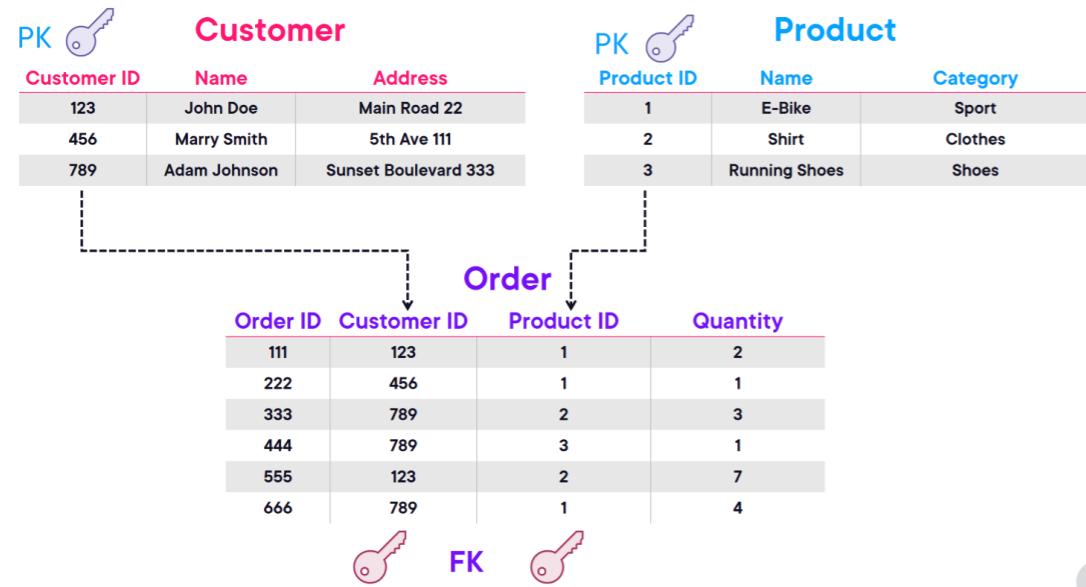
Tabela może zawierać dowolną liczbę wierszy

Klucz główny jednoznacznie identyfikuje wiersz w tabeli

Klucz obcy odwołuje się do wiersza w powiązanej tabeli



# Dane ustrukturyzowane



Wszystkie wiersze tabeli mają ten sam zestaw kolumn

Tabela może zawierać dowolną liczbę wierszy

Klucz główny jednoznacznie identyfikuje wiersz w tabeli

Klucz obcy odwołuje się do wiersza w powiązanej tabeli

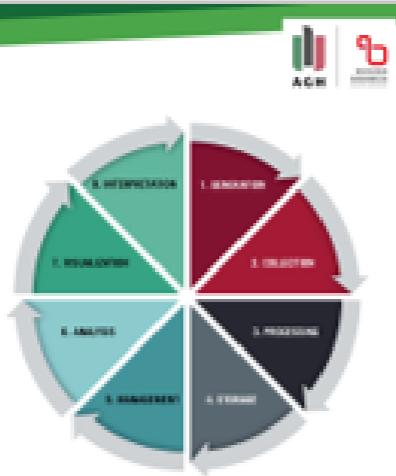
Zamodeluj

Example source:

Nikola Ilic

Data Mozart

@DataMozart | www.data-mozart.com



**Zarządzanie danymi**  
Zarządzanie danymi organizacyjnymi, generowanymi i pobieranymi

agh.edu.pl

**Zaspołeczeństwo:**

- Model hierarchiczny:** Stanowi hierarchyczny model, który reprezentuje strukturę danych organizacji, np. w postaci diagramu EER (Entity Relationship Diagram). Proces koncentruje się na relacjach i plikach, które reprezentują organizacyjne dane.
- Model hierarchiczny:** Oznacza organizacyjny język i skupia się na organizowaniu konsekwentnych, zdefiniowanych przez firmę informacji i danych organizacyjnych.
  - Diagram EER (Entity Relationship): Wykresów przedstawiających organizacyjne yapıły i ich relacje. Wykresy przedstawiają pojęcia i jednostki organizacyjne, a także ich relacje i powiązania.
  - Normalizacja EER: Proces normalizacji danych organizacyjnych, aby uniknąć niekonsekwencji (NA, TA, etc.).
  - Skrypty EML (Entity Modeling Language): Skrypty opisujące zmiany w organizacyjnych strukturach danych (NA, TA, etc.).
- Model logiczny:** Oznacza na poziomie konceptualnym struktury organizacyjne, które są łatwe do implementacji i łatwe do zmian.
  - Normalizacja EER: Proces normalizacji danych organizacyjnych, aby uniknąć niekonsekwencji (NA, TA, etc.).
  - Model logiczny: Wykresów przedstawiających organizacyjne yapıły i ich relacje.
- Model fizyczny:** Wykresów przedstawiających organizacyjne yapıły i ich relacje, które są łatwe do implementacji i łatwe do zmian.
  - Normalizacja EER: Proces normalizacji danych organizacyjnych, aby uniknąć niekonsekwencji (NA, TA, etc.).
  - Model fizyczny: Wykresów przedstawiających organizacyjne yapıły i ich relacje, które są łatwe do implementacji i łatwe do zmian.

agh.edu.pl

## Conceptual Data Model

## Logical Data Model

## Physical Data Model

## Purpose      Target Audience

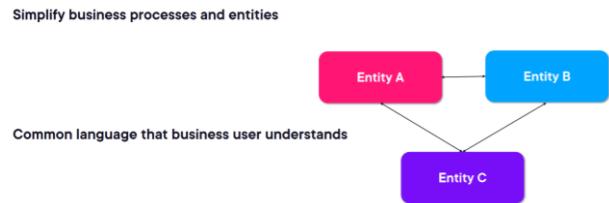
High-level overview	Business stakeholders, data professionals
Logical definition of data structures	Data engineers, data architects, data analysts
Low-level detail of physical data design	Database administrators, database developers

Example source:

Nikola Ilic  
Data Mozart

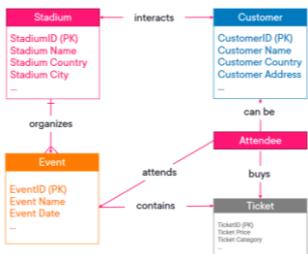
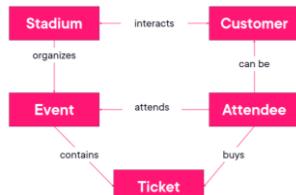
@DataMozart | www.data-mozart.com

# Konceptualnie



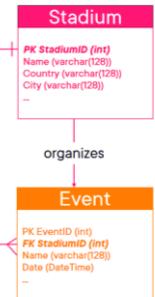
# Logicznie

- 1 Identify entity attributes
- 2 Identify candidate keys
- 3 Choose primary keys
- 4 Apply normalization/denormalization
- 5 Set relationships between entities
- 6 Identify the relationship cardinality
- 7 Iterate and fine-tune



# Fizycznie

- 1 Choose the platform
- 2 Translate logical entities into physical tables
- 3 Establish relationships
- 4 Apply normalization/denormalization
- 5 Apply table constraints
- 6 Create indexes and/or partitions
- 7 Extend with programmatic objects

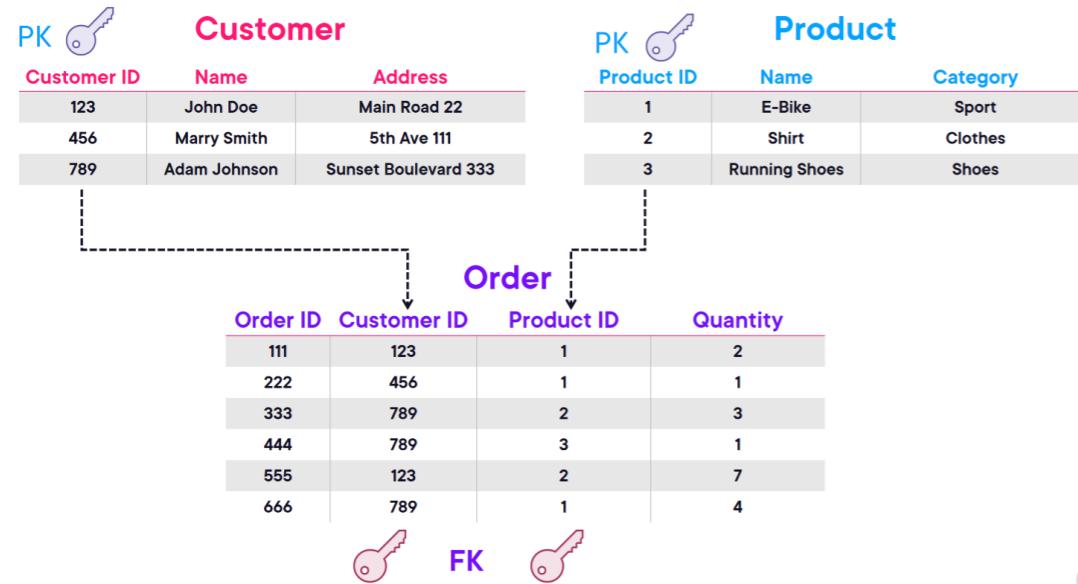


Example source:

Nikola Ilic  
Data Mozart

@DataMozart | www.data-mozart.com

## Dane ustrukturyzowane



Wszystkie wiersze tabeli mają ten sam zestaw kolumn

Tabela może zawierać dowolną liczbę wierszy

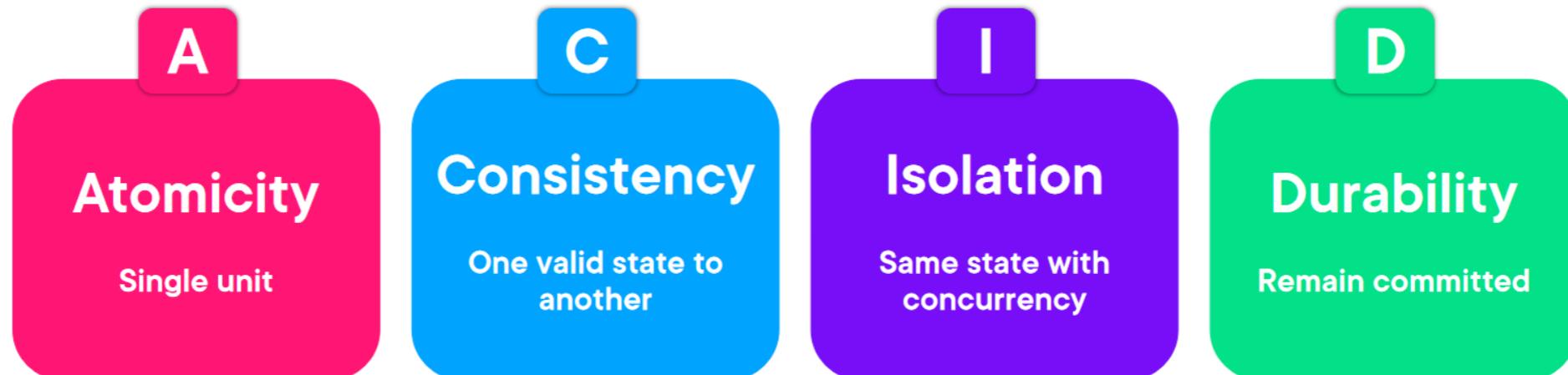
Klucz główny jednoznacznie identyfikuje wiersz w tabeli

Klucz obcy odwołuje się do wiersza w powiązanej tabeli

Relacyjne bazy danych są odpowiednie do przetwarzania transakcji

Relacyjne bazy danych są odpowiednie do przetwarzania transakcji

Cztery kluczowe właściwości definiują transakcje w relacyjnych bazach danych:  
**niepodzielność, spójność, izolacja i trwałość** — zwykle są one określane jako  
ACID



## Relacyjne bazy danych są odpowiednie do przetwarzania transakcji

**system przetwarzania danych polegający na wykonywaniu wielu transakcji odbywających się jednocześnie**, co ma miejsce na przykład w bankowości internetowej, w handlu detalicznym, podczas składania zamówień lub przy wysyłaniu wiadomości tekstowych

### OLTP

- Nadaje się do obsługi dużej liczby transakcji
- Wygodna obsługa instrukcji Insert, Update i Delete
- Wygodny do uruchamiania zapytań ad-hoc
- Koncentracja na integralności danych

## Relacyjne bazy danych są odpowiednie do przetwarzania transakcji

Zarządzanie danymi transakcyjnymi przy użyciu systemów komputerowych jest określane jako przetwarzanie transakcji online (OLTP). Systemy OLTP rejestrują interakcje biznesowe w miarę ich codziennego działania w organizacji i obsługują wykonywanie zapytań dotyczących tych danych w celu wnioskowania. System przetwarzania danych polegający na wykonywaniu wielu transakcji odbywających się jednocześnie, co ma miejsce na przykład w bankowości internetowej, w handlu detalicznym, podczas składania zamówień lub przy wysyłaniu wiadomości tekstowych itd.

### OLTP

- Nadaje się do obsługi dużej liczby transakcji
- Wygodna obsługa instrukcji Insert, Update i Delete
- Wygodny do uruchamiania zapytań ad-hoc
- Koncentracja na integralności danych

Systemy transakcyjne są przeznaczone do  
**WRITE**

Przetwarzanie transakcyjne -> WRITE

Przetwarzanie analityczne -> READ

Przetwarzanie analityczne online (OLAP) to technologia, która organizuje duże bazy danych biznesowych i obsługuje złożoną analizę. Może służyć do wykonywania złożonych zapytań analitycznych bez negatywnego wpływu na systemy transakcyjne.

Te bazy danych zwykle zawierają rekordy, które są wprowadzane pojedynczo. Często zawierają one wiele informacji, które są cenne dla organizacji. Bazy danych używane na potrzeby OLTP nie zostały jednak zaprojektowane do analizy. W związku z tym pobieranie odpowiedzi z tych baz danych jest kosztowne pod względem czasu i nakładu pracy. Systemy OLAP zostały zaprojektowane w celu ułatwienia wyodrębniania tych informacji analizy biznesowej z danych w bardzo wydajny sposób. Dzieje się tak, ponieważ bazy danych OLAP są zoptymalizowane pod kątem dużych obciążień odczytu i zapisu.

## OLTP

VS

## OLAP

- |  |  |
|--|--|
| <ul style="list-style-type: none"><li>1. Duża liczba małych transakcji</li><li>2. Operacje WRITE (Insert, Update)</li><li>3. Specyficzne dla branży (handel detaliczny, bankowość)</li><li>4. Transakcje jako źródło</li><li>5. Zwiększenie produktywności użytkowników końcowych</li><li>6. Znormalizowane bazy danych zapewniające wydajność</li></ul> | <ul style="list-style-type: none"><li>1. Duże ilości wolumenów danych</li><li>2. Operacje READ (Select)</li><li>3. Specyficzne dla przedmiotu (sprzedaż, marketing)</li><li>4. Zagregowane dane z transakcji</li><li>5. Zwiększenie produktywności analityków i kadry kierowniczej</li><li>6. <u>Zdenormalizowane</u> bazy danych do analizy</li></ul> |
|--|--|

## Przykładowe narzędzia

OLTP

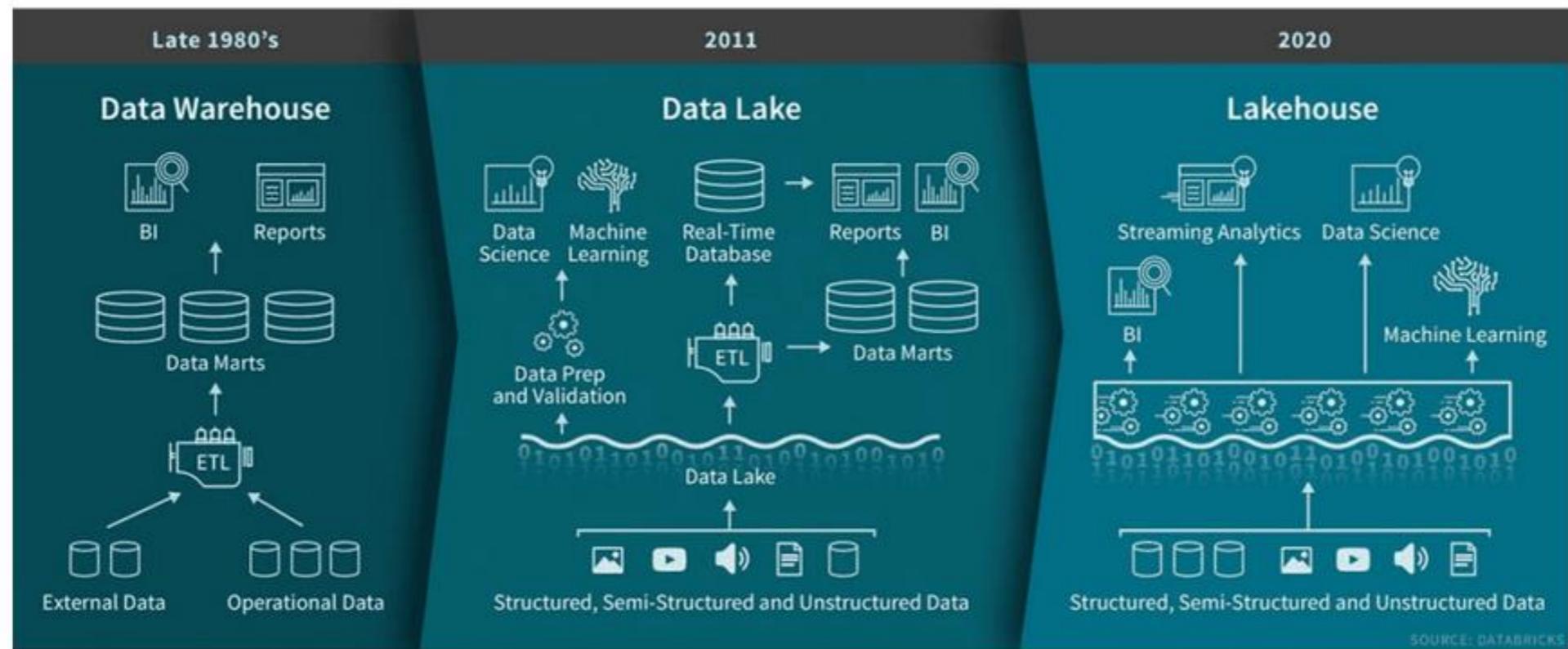
VS

OLAP

- Oracle Database
- Microsoft SQL Server
- MySQL
- PostgreSQL
- SQLite
- SAP HANA
- Amazon Aurora (MySQL, PostgreSQL)
- i wiele innych z ACID

- Microsoft SQL Server Analysis Services (SSAS)
- IBM Cognos
- Oracle OLAP
- SAP BusinessObjects Analysis
- MicroStrategy
- Pentaho Mondrian (Hitachi Vantara)
- Tableau Server
- Qlik Sense
- PowerBI (Connect to OLAP Cube)
- i wiele innych

Hybrid: Azure SQL Database, SQL Server in an Azure virtual machine, Azure Database for MySQL, Azure Database for PostgreSQL, Apache Cassandra with Spark, itd.



## Struktura

Magazyn danych (DWH) ma strukturę z góry określoną i uporządkowaną, zwykle zawiera przetworzone i gotowe do użycia dane, które są przygotowane do analizy biznesowej i raportowania.

Ustrukturyzowana

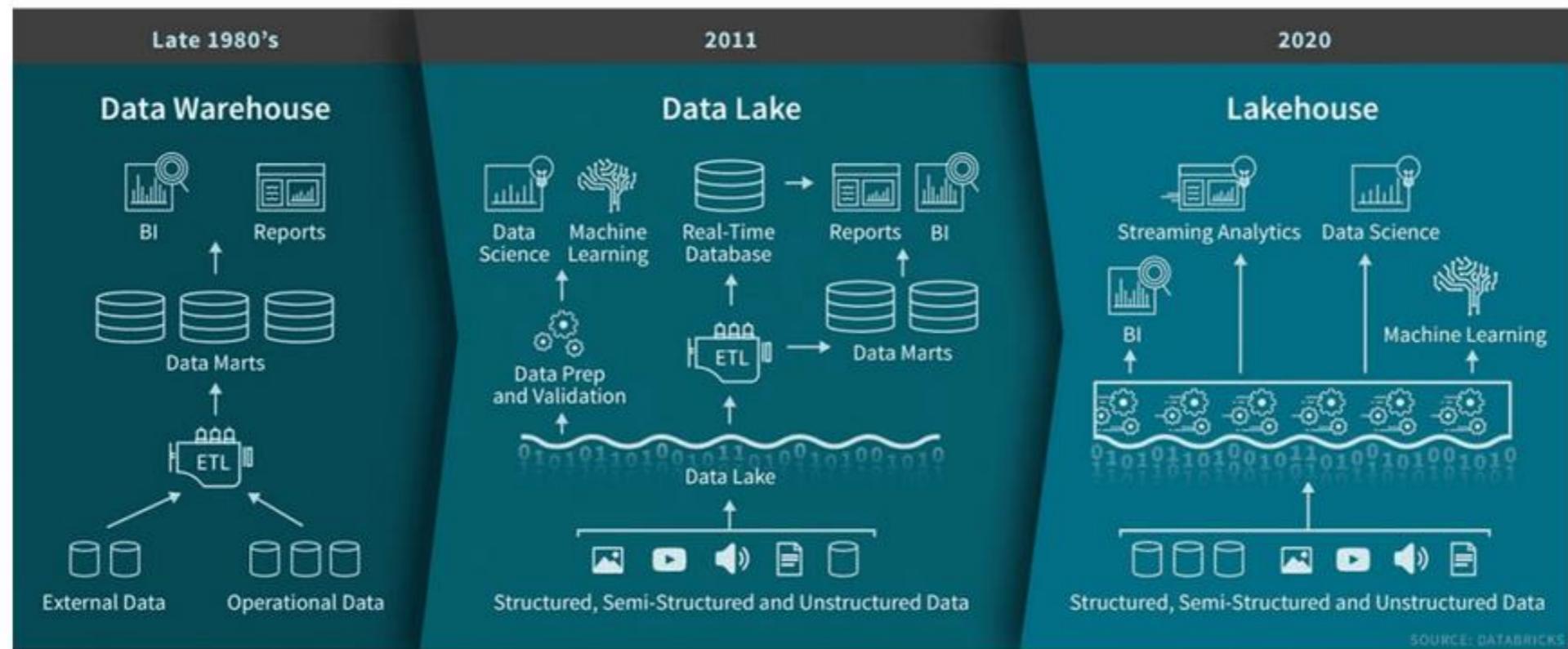
Jeziorko danych ma bardziej elastyczną strukturę, umożliwiając przechowywanie różnorodnych typów danych w oryginalnych lub surowych formatach, które nie wymagają wcześniejszego przetwarzania.

Nieustukturyzowana

Lakehouse danych łączy cechy magazynu danych (DWH) i jeziora danych, co oznacza, że ma strukturę gotową do użycia jak magazyn danych, ale może również przechowywać surowe lub przetworzone dane w oryginalnych formatach, jak jezioro danych.

Mieszana

<https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html>



## Architektura

Hurtownia danych jest zazwyczaj oparta na architekturze "top-down", co oznacza, że dane są starannie zintegrowane, przetworzone i skonsolidowane z różnych źródeł danych do centralnego repozytorium. Zazwyczaj wykorzystuje się modele wymiarowe lub relacyjne do przechowywania danych.

"top-down"

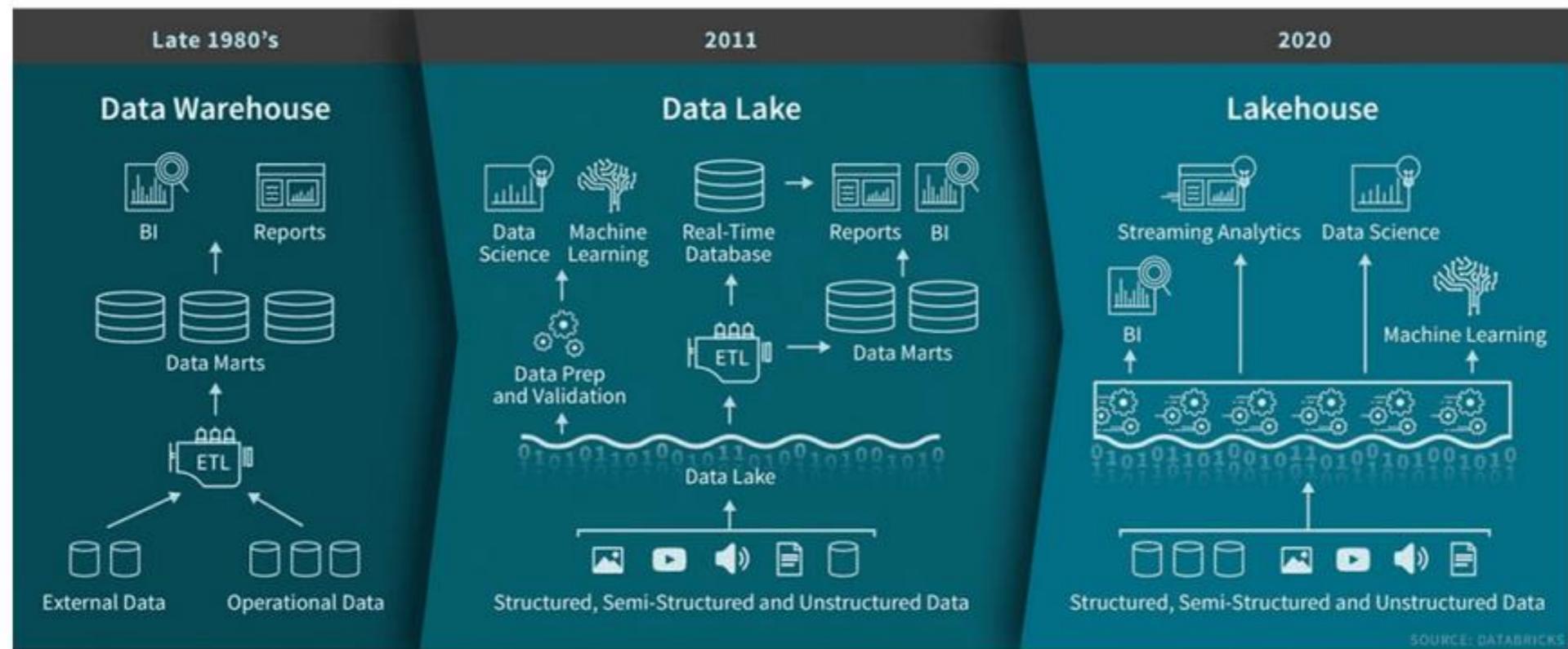
Jeziorko danych jest oparte na architekturze "bottom-up", gdzie dane są przechowywane w oryginalnej formie (surowe dane) bez wcześniejszej transformacji. Jeziorko danych może być zbudowane na różnych platformach, takich jak Hadoop, Amazon S3, czy Azure Data Lake Storage.

bottom-up

Lakehouse danych jest koncepcją, która łączy cechy hurtowni danych i jeziora danych. W Lakehouse danych, dane są przechowywane w jeziorze danych w ich oryginalnej formie, ale również są dostępne w formie przetworzonej i zoptymalizowanej do analizy.

mieszana

<https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html>



## Model

W hurtowni danych stosuje się zazwyczaj modele wymiarowe, takie jak model gwiazdy lub model płatka śniegu, które umożliwiają analizę danych w sposób zorientowany na temat (ang. subject-oriented)

Jeziorko danych może przechowywać różnorodne typy danych, w tym dane strukturalne, półstrukturalne i niestrukturalne. Nie ma konieczności stosowania ściśle zdefiniowanych schematów danych

Lakehouse danych może wykorzystywać modele danych charakterystyczne dla zarówno hurtowni danych, jak i jeziora danych. Może to obejmować zarówno modele wymiarowe, jak i przechowywanie surowych danych w celu późniejszej analizy

<https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html>


 Q  

Learn about the over 450,000  
Kimball Toolkits sold

[Home](#) [Resources](#) [About](#) [Contact](#)

## Dimensional Modeling Techniques

[Home](#) / [Data Warehouse and Business Intelligence Resources](#) / [Kimball Techniques](#) / Dimensional Modeling Techniques

Ralph Kimball introduced the data warehouse/business intelligence industry to dimensional modeling in 1996 with his seminal book, *The Data Warehouse Toolkit*. Since then, the Kimball Group has extended the portfolio of best practices.

Drawn from *The Data Warehouse Toolkit, Third Edition*, the "official" Kimball dimensional modeling techniques are described on the following links and attached .pdf:

### Fundamental Concepts

- Gather business requirements and data realities
- Collaborative dimensional modeling workshops
- Four step dimensional design process
- Business processes
- Grain
- Dimensions for descriptive context
- Facts for measurements
- Star schemas and OLAP cubes
- Graceful extensions to dimensional models

### Basic Fact Table Techniques

- Fact table structure
- Additive, semi-additive, and non-additive facts
- Nulls in fact tables
- Conformed facts



Bob Becker, Margy Ross, Warren Thornthwaite  
Joy Mundy, Ralph Kimball, Julie Kimball

## Ralph Kimball



**"The Data Warehouse Toolkit" (1996)**

**Dimensional modeling "Bible"**

**"Bottom-up" approach**

- Identify and model KEY business processes

<https://www.kimballgroup.com/author/ralph/>

## Proces Projektowania Modelu Wymiarowego Kimballa (4 Kroki)

Proces projektowania modelu wymiarowego Kimballa to powszechnie stosowana metoda tworzenia hurtowni danych dla potrzeb Business Intelligence (BI) i analizy. Skupia się on na tworzeniu schematu gwiazdzistego, specyficznej struktury, w której centralną tabelę faktów otaczają tabele wymiarowe. Oto szczegółowe omówienie 4-etapowego procesu:

### 1. Wybór Procesu Biznesowego:

Ten początkowy krok polega na identyfikacji konkretnego procesu biznesowego, który chcesz analizować. Może to być sprzedaż, kampanie marketingowe, zachowania klientów lub dowolna inna podstawowa funkcja w organizacji.

### 2. Określenie Granularności:

Granularność określa poziom szczegółowości przechowywania danych w tabeli faktów. Definiuje ona najniższy poziom agregacji dla Twojej analizy. Typowe opcje granularności obejmują transakcje, dni, tygodnie, miesiące lub kwartały. Wybór odpowiedniej granularności zależy od procesu biznesowego i typów analiz, które chcesz przeprowadzić.

### 3. Identyfikacja Wymiarów:

Wymiary dostarczają kontekstu i opisowych atrybutów dla faktów przechowywanych w tabeli faktów. Odpowiadają na pytania „kto, co, kiedy, gdzie, dlaczego i jak” związane z procesem biznesowym. Przykłady wymiarów to klient, produkt, czas, lokalizacja lub kanał. Każda tabela wymiarowa powinna mieć unikalny identyfikator (klucz podstawowy) i odpowiednie atrybuty opisujące ten wymiar.

### 4. Identyfikacja Faktów:

Fakty to wartości ilościowe przechowywane w tabeli faktów. Są to wartości liczbowe, które chcesz analizować, takie jak kwota sprzedaży, sprzedana ilość, współczynnik kliknięć lub wartość klienta w ciągu całego życia. Fakty są zazwyczaj przechowywane jako klucze obce łączące się z tabelami wymiarowymi, co pozwala na segmentację i analizowanie danych na podstawie różnych wymiarów.

Przestrzeganie tych kroków pomaga zapewnić dobrze zorganizowaną hurtownię danych, ułatwiającą wydajne wyszukiwanie i analizę danych dla celów BI. Istnieją dodatkowe kroki, które mogą być zaangażowane w kompletny projekt modelu wymiarowego, takie jak definiowanie relacji między tabelami, ustawianie typów danych i ograniczeń oraz dokumentowanie modelu.

## Proces Projektowania Modelu Wymiarowego Kimballa (4 Kroki)

Proces projektowania modelu wymiarowego Kimballa to powszechnie stosowana metoda tworzenia hurtowni danych dla potrzeb Business Intelligence (BI) i analizy. Skupia się on na tworzeniu schematu gwiaździstego, specyficznej struktury, w której centralną tabelę faktów otaczają tabele wymiarowe. Oto szczegółowe omówienie 4-etapowego procesu:

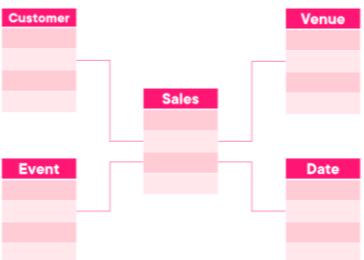
### 1 Selecting the business process

Event	Venue	Customer	Quantity	Amount	Employee	Ticket Type	Country	Date
Barca vs Man Utd	Camp Nou	John Doe	2	200€	Mike	VIP	Spain	20220501
Man City vs Man Utd	Etihad Stadium	Melanie D.	3	300	Nikola	Regular	UK	20220501
Liverpool vs Chelsea	Emirates	Greg H.	1	500	Mark	Regular	UK	20220501
Man City vs Liverpool	Broadway	Pascal G.	4	400€	Mark	VIP	USA	20220501
Chelsea vs Liverpool	Broadway	Frankie G.	2	200€	Nikola	VIP	USA	20220501

### 2 Declare the grain

The lowest level of detail captured by the business process

Event	Venue	Customer	Quantity	Amount	Employee	Ticket Type	Country	Date
Barca vs Man Utd	Camp Nou	John Doe	2	200€	Mike	VIP	Spain	20220501
Man City vs Man Utd	Etihad Stadium	Melanie D.	3	300	Nikola	Regular	UK	20220501
Liverpool vs Chelsea	Emirates	Greg H.	1	500	Mark	Regular	UK	20220501
Man City vs Liverpool	Broadway	Pascal G.	4	400€	Mark	VIP	USA	20220501
Chelsea vs Liverpool	Broadway	Frankie G.	2	200€	Nikola	VIP	USA	20220501



Model Gwiazdy (Star schema)

### 3 Identify the dimensions



Dimension

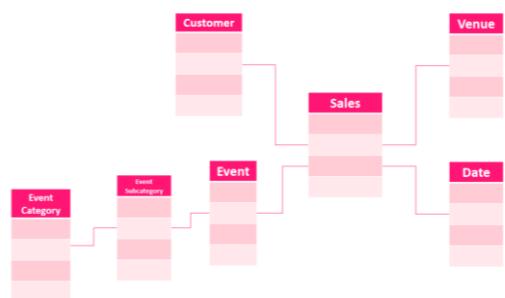


Lookup table

- Who?
- What?
- When?
- Where?
- Why?

W

When did we sell a ticket?  
Where did we sell a ticket?  
What type of ticket did we sell?



Płatek śniegu (Snowflake schema)

### 4 Identify the facts

- 1 In most cases numeric values
- 2
- 3

How many tickets did we sell?  
How much revenue did we make?



Normalizacja to proces organizowania danych w bazie danych. Skutkuje modelami (np. płatka śniegu).

**Normalizacja** to proces organizowania danych w bazie danych w celu wyeliminowania **redundancji i anomalii**. Celem normalizacji jest stworzenie bazy danych, która jest:

- **Spójna:** Dane w różnych tabelach są ze sobą zgodne i odnoszą się do siebie w logiczny sposób.
- **Efektywna:** Operacje na bazie danych, takie jak wyszukiwanie, dodawanie i usuwanie danych, są wykonywane szybko i sprawnie.
- **Łatwa w utrzymaniu:** Baza danych jest łatwa do aktualizacji i modyfikowania bez wprowadzania błędów.

Najczęstsze poziomy normalizacji to:

- **Pierwsza postać normalna (1NF):**  
Eliminuje powtarzające się wartości w kolumnach.
- 
- **Druga postać normalna (2NF):**  
Eliminuje redundancję wynikającą z zależności między kolumnami.
- **Trzecia postać normalna (3NF):**  
Eliminuje redundancję wynikającą z zależności przechodnich między kolumnami.

NrPozyji	NumerZam...	NazwaKlienta	Adres	KodPocztowy	Miasto	Województwo	DataZamowienia	ElementZamowienia	Ilosc	CenaJedn	WartZamNetto	Vat	WartZamBrutto
1	101	Jan Kowalski	ul. Jana Pawła 12	61-600	Poznań	Wielkopolskie	2012-01-02 00:00:00	Opony 205 R16	4	300,00	1200,00	23	1476,00
2	102	Anna Dymna	ul. Staszica 1	30-600	Kraków	Małopolskie	2012-03-22 00:00:00	Alufelgi Silver	4	550,00	2200,00	23	2706,00
3	103	Piotr Wawrzyniak	al. Niepodległości 1	30-600	Kraków	Małopolskie	2012-03-22 00:00:00	Alufelgi Silver	4	550,00	2200,00	23	2706,00
4	104	Jan Kowalski	ul. Jana Pawła 12	61-600	Poznań	Wielkopolskie	2012-10-22 00:00:00	Komplet żarówek	1	80,00	80,00	23	98,40
5	105	Jan Kowalski	ul. Poznańska 8	21-120	Wrocław	Dolnośląskie	2012-05-22 00:00:00	Plyn do spryskiwacza	1	10,00	15,00	23	18,45
6	105	Jan Kowalski	ul. Poznańska 8	21-120	Wrocław	Dolnośląskie	2012-05-22 00:00:00	Trójkąt ostrzegawczy	1	5,00	15,00	23	18,45

Klient							Detaile zamówienia						
NrPozyji	NumerZam...	NazwaKlienta	Adres	KodPocztowy	Miasto	Województwo	DataZamowienia	ElementZamowienia	Ilosc	CenaJedn	WartZamNetto	Vat	WartZamBrutto
1	101	Jan Kowalski	ul. Jana Pawła 12	61-600	Poznań	Wielkopolskie	2012-01-02 00:00:00	Opony 205 R16	4	300,00	1200,00	23	1476,00
2	102	Anna Dymna	ul. Staszica 1	30-600	Kraków	Małopolskie	2012-03-22 00:00:00	Alufelgi Silver	4	550,00	2200,00	23	2706,00
3	103	Piotr Wawrzyniak	al. Niepodległości 1	30-600	Kraków	Małopolskie	2012-03-22 00:00:00	Alufelgi Silver	4	550,00	2200,00	23	2706,00
4	104	Jan Kowalski	ul. Jana Pawła 12	61-600	Poznań	Wielkopolskie	2012-10-22 00:00:00	Komplet żarówek	1	80,00	80,00	23	98,40
5	105	Jan Kowalski	ul. Poznańska 8	21-120	Wrocław	Dolnośląskie	2012-05-22 00:00:00	Plyn do spryskiwacza	1	10,00	15,00	23	18,45
6	105	Jan Kowalski	ul. Poznańska 8	21-120	Wrocław	Dolnośląskie	2012-05-22 00:00:00	Trójkąt ostrzegawczy	1	5,00	15,00	23	18,45

NumerZamowienia	IDKlient	DataZamowienia	WartZamNetto	WartZamBrutto	Vat %
101	1	2012-01-02 00:00:00	1200,00	1476,00	23,00
102	2	2012-03-22 00:00:00	2200,00	2706,00	23,00
103	3	2012-03-22 00:00:00	2200,00	2706,00	23,00
104	1	2012-10-22 00:00:00	80,00	98,40	23,00
105	4	2012-05-22 00:00:00	15,00	18,45	23,00

Najczęstsze poziomy normalizacji to:

- **Pierwsza postać normalna (1NF):**  
Eliminuje powtarzające się wartości w kolumnach.
- 
- **Druga postać normalna (2NF):**  
Eliminuje redundancję wynikającą z zależności między kolumnami.
- **Trzecia postać normalna (3NF):**  
Eliminuje redundancję wynikającą z zależności przechodnich między kolumnami.

**Employee**

Employee ID	Employee Name	Job ID	Job Name	State ID	State
1	John Doe	1	Teacher	1	WA
1	John Doe	2	Director	1	WA
2	Marry Poppins	1	Teacher	2	NY
2	Marry Poppins	2	Director	2	NY
3	Lady Bug	2	Director	2	NY

**Employee**

Employee ID	Emp Name	State ID	State
1	John Doe	1	WA
2	Marry Poppins	2	NY
3	Lady Bug	2	NY

**Employee Job**

Employee ID	Job ID
1	1
1	2
2	1
2	1
3	2

**Job**

Job ID	Job Name
1	Teacher
2	Director

**Employee**

Employee ID	Emp Name	State ID	State
1	John Doe	1	WA
2	Marry Poppins	2	NY
3	Lady Bug	2	NY

**Employee Job**

Employee ID	Job ID
1	1
1	2
2	1
2	1
3	2

**Job**

Job ID	Job Name
1	Teacher
2	Director

**State**

State ID	State Name
1	WA
2	NY

### Korzyści z normalizacji:

- **Zmniejszenie redundancji:** Prowadzi to do mniejszego rozmiaru bazy danych i oszczędza miejsce na dysku.
- **Zwiększenie spójności:** Dane w różnych tabelach są ze sobą zgodne, co zmniejsza ryzyko błędów.
- **Ułatwienie aktualizacji:** Łatwiej jest aktualizować dane bez wprowadzania błędów.
- **Poprawa wydajności:** Operacje na bazie danych, takie jak wyszukiwanie, dodawanie i usuwanie danych, mogą być wykonywane szybciej.

### Wady normalizacji:

- **Zwiększona złożoność:** Normalizacja może uczynić strukturę bazy danych bardziej złożoną i trudniejszą do zrozumienia.
- **Spadek wydajności:** W niektórych przypadkach normalizacja może spowolnić operacje na bazie danych.
- **Wymagania dotyczące dodatkowych zapytań:** Dostęp do niektórych danych może wymagać wykonania większej liczby zapytań.

### Korzyści z normalizacji:

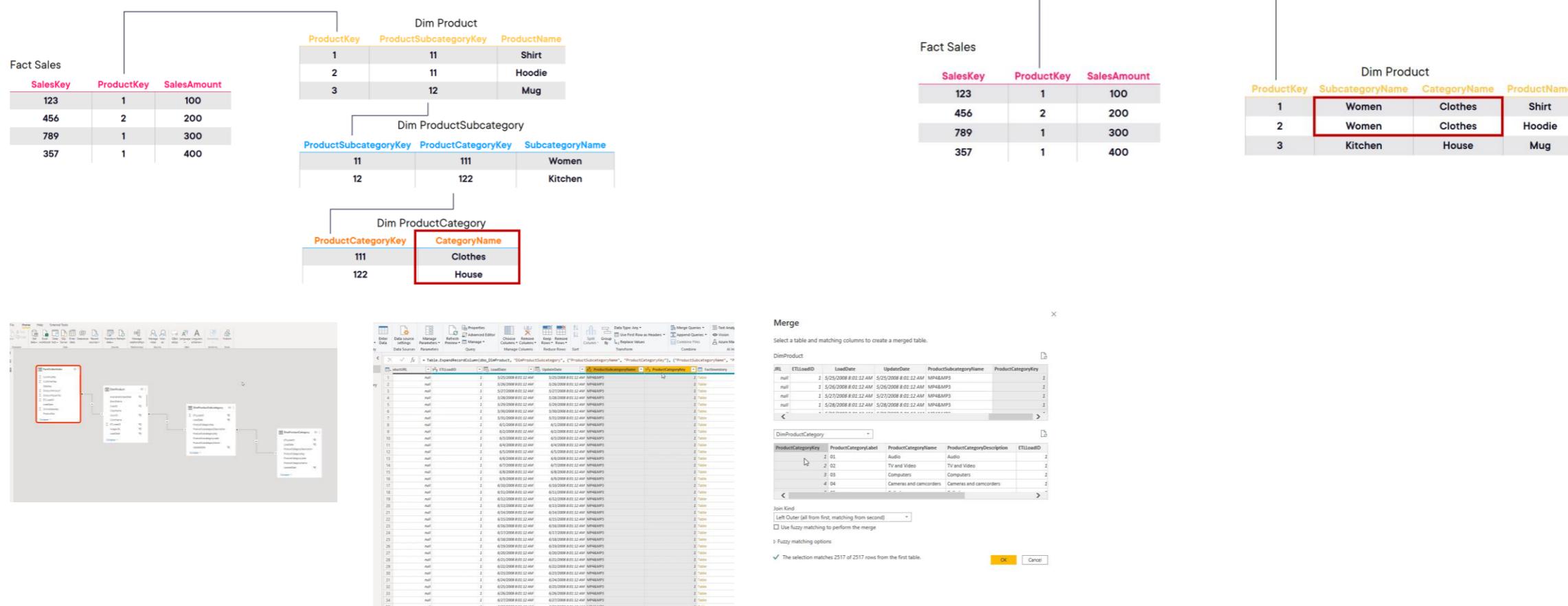
- **Zmniejszenie redundancji:** Prowadzi to do mniejszego rozmiaru bazy danych i oszczędza miejsce na dysku.
- **Zwiększenie spójności:** Dane w różnych tabelach są ze sobą zgodne, co zmniejsza ryzyko błędów.
- **Ułatwienie aktualizacji:** Łatwiej jest aktualizować dane bez wprowadzania błędów.
- **Poprawa wydajności:** Operacje na bazie danych, takie jak wyszukiwanie, dodawanie i usuwanie danych, mogą być wykonywane szybciej.

### Wady normalizacji:

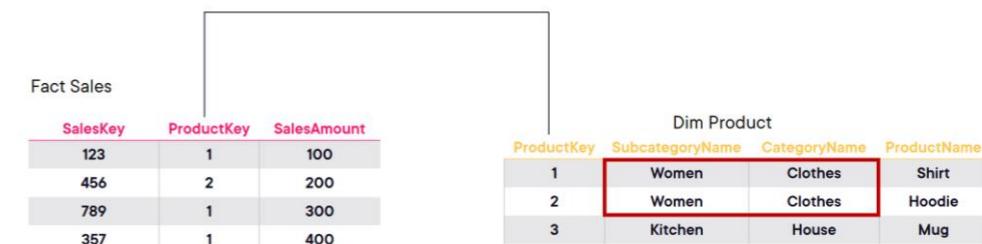
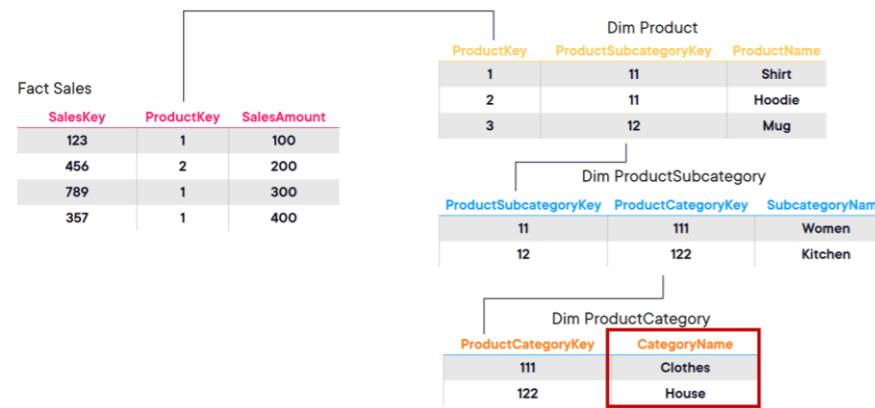
- **Zwiększona złożoność:** Normalizacja może uczynić strukturę bazy danych bardziej złożoną i trudniejszą do zrozumienia.
- **Spadek wydajności:** W niektórych przypadkach normalizacja może spowolnić operacje na bazie danych.
- **Wymagania dotyczące dodatkowych zapytań:** Dostęp do niektórych danych może wymagać wykonania większej liczby zapytań.

**Denormalizacja** jest procesem **wprowadzania kontrolowanej redundancji** do bazy danych w celu **optymalizacji wydajności zapytań**. Oznacza to, że w celu przyspieszenia dostępu do danych i zmniejszenia liczby operacji dołączania tabel, celowo powtarzane są pewne dane w różnych tabelach.

**Denormalizacja** jest procesem **wprowadzania kontrolowanej redundancji** do bazy danych w celu **optymalizacji wydajności zapytań**. Oznacza to, że w celu przyspieszenia dostępu do danych i zmniejszenia liczby operacji dołączania tabel, celowo powtarzane są pewne dane w różnych tabelach.



**Denormalizacja** jest procesem **wprowadzania kontrolowanej redundancji** do bazy danych w celu **optymalizacji wydajności zapytań**. Oznacza to, że w celu przyspieszenia dostępu do danych i zmniejszenia liczby operacji dołączania tabel, celowo powtarzane są pewne dane w różnych tabelach.



### Kiedy ma sens ?

- **Częste zapytania dotyczące powiązanych danych:** Jeśli często wykonujesz zapytania, które wymagają dołączenia danych z wielu tabel, denormalizacja może przyspieszyć te zapytania, duplikując dane w tabeli, do której najczęściej wykonujesz zapytania.
- **Złożone operacje dołączania:** Jeśli Twoje zapytania wymagają złożonych operacji dołączania tabel, denormalizacja może je uprościć i przyspieszyć.

### Wady denormalizacji:

- **Zwiększona redundancja:** Prowadzi to do większego rozmiaru bazy danych i wymaga więcej miejsca na dysku.
- **Zmniejszona spójność:** Dane w różnych tabelach mogą nie być ze sobą zgodne, co może prowadzić do błędów.
- **Trudniejsza aktualizacja:** Aktualizacja danych może być bardziej skomplikowana i czasochłonna ze względu na redundancję.
- **Zwiększona złożoność:** Struktura bazy danych może stać się bardziej złożona i trudniejsza do zrozumienia.

## Czym jest architektura medalionowa?

Architektura medalionowa to wzorzec projektowania dla Data Lakehouse, który opisuje serię warstw danych, zapewniających jakość i spójność przechowywanych danych. Nazwa pochodzi od kształtu medalionu, który symbolizuje centralną warstwę metadanych, otoczoną wieloma warstwami danych surowych i przetworzonych.



### Warstwy architektury medalionowej:

- Warstwa surowa (Raw):** Zawiera nieprzetworzone dane w ich oryginalnym formacie, takie jak pliki CSV, JSON lub XML.
- Warstwa filtrowania (Curated):** Dane z warstwy surowej są filtrowane, oczyszczane i weryfikowane pod kątem błędów i nieścisłości.
- Warstwa wzbogacona (Enriched):** Dane z warstwy filtrowania są wzbogacane o dodatkowe informacje z zewnętrznych źródeł, takie jak dane referencyjne lub metadane.
- Warstwa tematyczna (Thematic):** Dane z warstwy wzbogaconej są zorganizowane w tematyczne obszary zainteresowania, ułatwiając dostęp i analizę.
- Warstwa medalionu (Medallion):** Zawiera metadane opisujące wszystkie warstwy danych, ich strukturę, pochodzenie i zastosowanie.

### Zalety architektury medalionowej:

- Niepodzielność:** Zapewnia spójność i wiarygodność danych w całej architekturze.
- Spójność:** Gwarantuje, że wszystkie warstwy danych są ze sobą zgodne i odnoszą się do tego samego źródła.
- Izolacja:** Pozwala na niezależne przetwarzanie i aktualizowanie danych w każdej warstwie bez wpływu na inne warstwy.
- Trwałość:** Zapewnia długoterminowe przechowywanie i dostęp do danych.

### Zastosowania architektury medalionowej:

- Analiza danych:** Ułatwia analizę danych z różnych źródeł i generowanie spójnych wyników.
- Nauka maszynowa:** Zapewnia wysokiej jakości dane do trenowania modeli uczenia maszynowego.
- Raporty i wizualizacje:** Umożliwia tworzenie dokładnych i aktualnych raportów i wizualizacji danych.

### Przykładowe narzędzia do implementacji architektury medalionowej:

- Databricks:** <https://learn.microsoft.com/en-us/azure/databricks/lakehouse/>
- Amazon S3 i AWS Glue:** <https://aws.amazon.com/s3/>
- Google Cloud Storage i BigQuery:** <https://cloud.google.com/bigquery>

### Dodatkowe informacje:

- Więcej o architekturze medalionowej w Azure Databricks: <https://learn.microsoft.com/en-us/fabric/onelake/onelake-medallion-lakehouse-architecture>
- [Scholar](#)

## Lakehouse zones



- ✓ Land data from external sources in its original state
- ✓ Serve as a repository of the historical archive of source data
- ✓ Contains unvalidated data

✓ Stores the data in Parquet/Delta

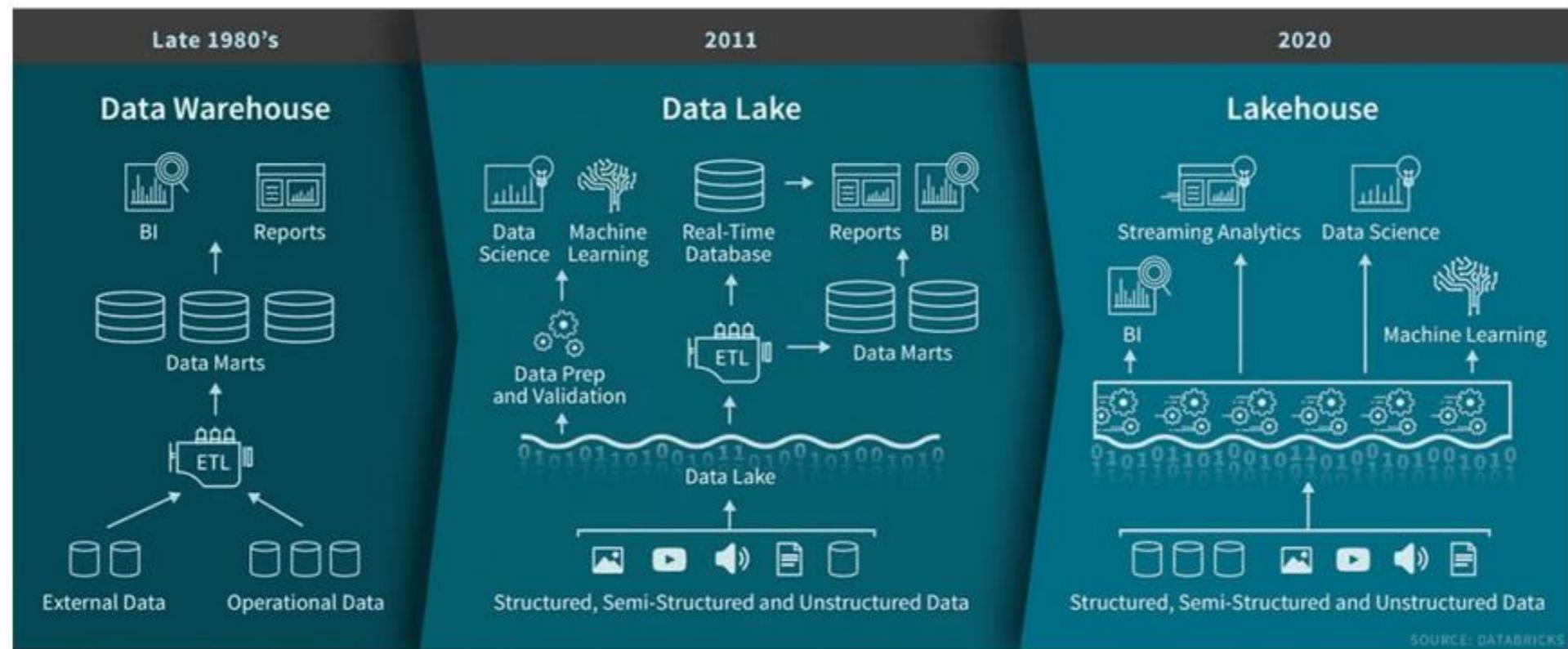


- ✓ Conformed and cleaned data from the bronze layer
- ✓ Ad-hoc analysis, machine learning workloads
- ✓ Contains enriched and validated data
- ✓ Data model normalized to a 3<sup>rd</sup> normal form
- ✓ Stores the data in Delta/Parquet



- ✓ Structured and organized data for specific project requirements
- ✓ Data additionally cleaned and refined
- ✓ Complex business logic and specific calculations
- ✓ Data model is a Kimball-style star schema
- ✓ Stores the data preferably in Delta, alternatively in Parquet

<http://www.unstructureddatatips.com/what-is-data-lakehouse/>



### Typy danych

Przechowywane są strukturalne dane biznesowe, które są skonsolidowane, zintegrowane i poddane transformacji, aby zapewnić spójność i jakość danych

Jezioro danych przechowuje różnorodne dane, w tym dane surowe, dane przetworzone, dane w czasie rzeczywistym oraz dane zewnętrzne

Lakehouse danych przechowuje zarówno surowe, jak i przetworzone dane. Dostępne są również różnorodne typy danych, podobnie jak w jeziorze danych

<https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html>

Dwa popularne formaty plików, które zyskały znaczną uwagę, to pliki w formacie Parquet i Delta. Podczas gdy Parquet jest kolumnowym formatem przechowywania danych znany z wysokiej kompresji i wydajności zapytań, pliki w formacie Delta zapewniają możliwości transakcyjne i zgodność z ACID.

### Parquet Files

Pliki Parquet Pliki Parquet to kolumnowy format pamięci masowej.

Row-Based Storage

Product	Customer	Country	Date	Sales Amount
Ball	John Doe	USA	2023-01-01	100
T-Shirt	John Doe	USA	2023-01-02	200
Socks	Maria Adams	UK	2023-01-01	300
Socks	Antonio Grant	USA	2023-01-03	100
T-Shirt	Maria Adams	UK	2023-01-02	500
Socks	John Doe	USA	2023-01-05	200

Row-Based Storage

Product	Customer	Country	Date	Sales Amount
Ball		USA	2023-01-01	100
T-Shirt		USA	2023-01-02	200
Socks		UK	2023-01-01	300
Socks		USA	2023-01-03	100
T-Shirt		UK	2023-01-02	500
Socks		USA	2023-01-05	200

Column-Based Storage

Column 1	Column 2	Column 3	Column 4	Column 5
Product	Customer	Country	Date	Sales Amount
Ball	John Doe	USA	2023-01-01	100
T-Shirt	John Doe	USA	2023-01-02	200
Socks	Maria Adams	UK	2023-01-01	300
Socks	Antonio Grant	USA	2023-01-03	100
T-Shirt	Maria Adams	UK	2023-01-02	500
Socks	John Doe	USA	2023-01-05	200

Parquet Storage

Column 1	Column 2	Column 3	Column 4	Column 5
Product	Customer	Country	Date	Sales Amount
Row group 1	Ball	John Doe	USA	2023-01-01
	T-Shirt	John Doe	USA	2023-01-02
Row group 2	Socks	Maria Adams	UK	2023-01-01
	Socks	Antonio Grant	USA	2023-01-03
Row group 3	T-Shirt	Maria Adams	UK	2023-01-02
	Socks	John Doe	USA	2023-01-05

Projection and Predicate(s)

Projection = SELECT	Predicate(s) = WHERE
Column 1 Column 2 Column 3 Column 4 Column 5	
Product Customer Country Date Sales Amount	
Row group 1 Ball John Doe USA 2023-01-01 100	
Row group 2 T-Shirt John Doe USA 2023-01-02 200	
Row group 3 Socks Maria Adams UK 2023-01-01 300	
Row group 4 Socks Antonio Grant USA 2023-01-03 100	
Row group 5 T-Shirt Maria Adams UK 2023-01-02 500	
Row group 6 Socks John Doe USA 2023-01-05 200	

The engine will skip scanning these records!

### Delta Format Files

Pliki w formacie delta to rodzaj technologii przechowywania plików, która służy do przechowywania modyfikacji plików zamiast całych zmienionych plików.

Row-Based Storage: 5 Columns + 6 Rows

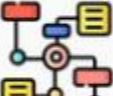
Product	Customer	Country	Date	Sales Amount
Row 1	Ball	John Doe	USA	2023-01-01
Row 2	T-Shirt	John Doe	USA	2023-01-02
Row 3	Socks	Maria Adams	UK	2023-01-01
Row 4	Socks	Antonio Grant	USA	2023-01-03
Row 5	T-Shirt	Maria Adams	UK	2023-01-02
Row 6	Socks	John Doe	USA	2023-01-05

Column-Based Storage: 2 Columns + 6 Rows

Column 1	Column 2	Column 3	Column 4	Column 5
Product	Customer	Country	Date	Sales Amount
Row 1	Ball	John Doe	USA	2023-01-01
Row 2	T-Shirt	John Doe	USA	2023-01-02
Row 3	Socks	Maria Adams	UK	2023-01-01
Row 4	Socks	Antonio Grant	USA	2023-01-03
Row 5	T-Shirt	Maria Adams	UK	2023-01-02
Row 6	Socks	John Doe	USA	2023-01-05

	 PARQUET	 DELTA
	<b>STORAGE FORMAT</b>	Uses columnar storage format which allows for efficient compression and querying of specific columns from the files.
	<b>COMPRESSION</b>	Provides strong compression capabilities for columnar data through techniques like run-length encoding and dictionary encoding. This allows for better storage utilisation.
	<b>QUERY PERFORMANCE</b>	Faster for selective queries that filter on a few columns since it can skip reading entire rows/blocks. But slower for queries that require reading all columns.
	<b>SCHEMA EVOLUTION</b>	Requires writing all schema changes to new files. Old data remains untouched but queries need to handle multiple schemas.
	<b>ACID TRANSACTIONS</b>	Does not support transactions. Not suitable for applications requiring strong consistency guarantees.
	<b>METADATA STORAGE</b>	Metadata is stored in each data file. Requires scanning all files to get overall metadata.

Graphic: <https://www.linkedin.com/pulse/should-i-use-parquet-files-delta-format-comparative-analysis-s/>

	 PARQUET	 DELTA
 ANALYTICS WORKLOADS	Analytics queries that filter on certain columns are highly optimised. This is useful for huge datasets analysed via batch analytics processes.	When compared to Parquet, it may be less optimised for extremely selective searches on huge datasets.
 TRANSACTIONAL WORKLOADS	Not suitable as it doesn't support ACID transactions required for transactional systems.	Specifically designed for workloads with mixed transactional and analytical processing. Supports ACID properties.
 ONLINE TRANSACTION PROCESSING	Not a good fit due to lack of transactions and poorer performance for full table scans.	Better suited as it delivers row-level ACID transactions and scalable ingest with Spark SQL. Supports OLTP workloads.
 EVOLVING/CHANGING SCHEMAS	Requires writing schema changes to new files. Complex for schemas that change frequently.	Explicit schema support and versioning makes it easier to evolve schemas without dataset rewrites.
 SMALL/MEDIUM SIZED DATASETS	Overhead of columnar storage may not be worthwhile for small-medium sized data.	Set oriented operations and ACID make it more user friendly for modest sized evolving datasets.
 REAL-TIME APPLICATIONS	Not suitable due to lack of transactions and overhead of merging columnar data.	Lower latency ACID transactions and simpler metadata makes it more viable for real-time or low latency apps.

Graphic: <https://www.linkedin.com/pulse/should-i-use-parquet-files-delta-format-comparative-analysis-s/>

## Ale warto skonfrontować:

- <https://delta.io/blog/delta-lake-vs-parquet-comparison/>
- Schneider, Jan, et al. "Assessing the Lakehouse: Analysis, Requirements and Definition." *ICEIS* (1). 2023.

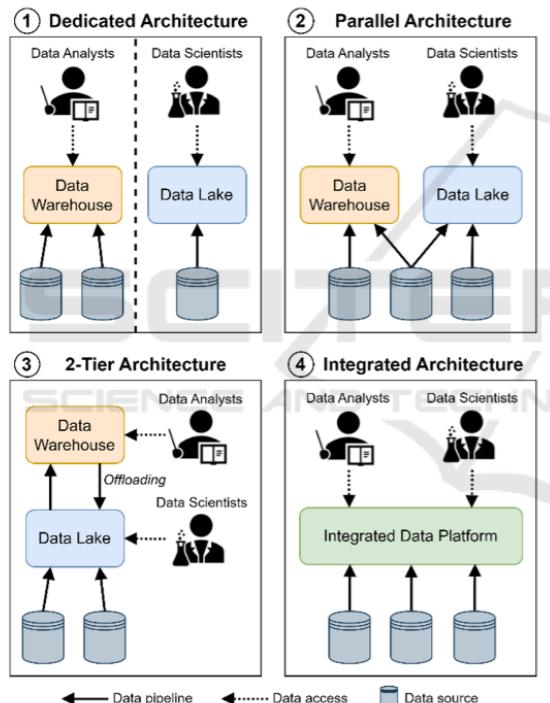


Figure 1: Integration patterns for combining data warehouses and data lakes in enterprise analytics architectures.

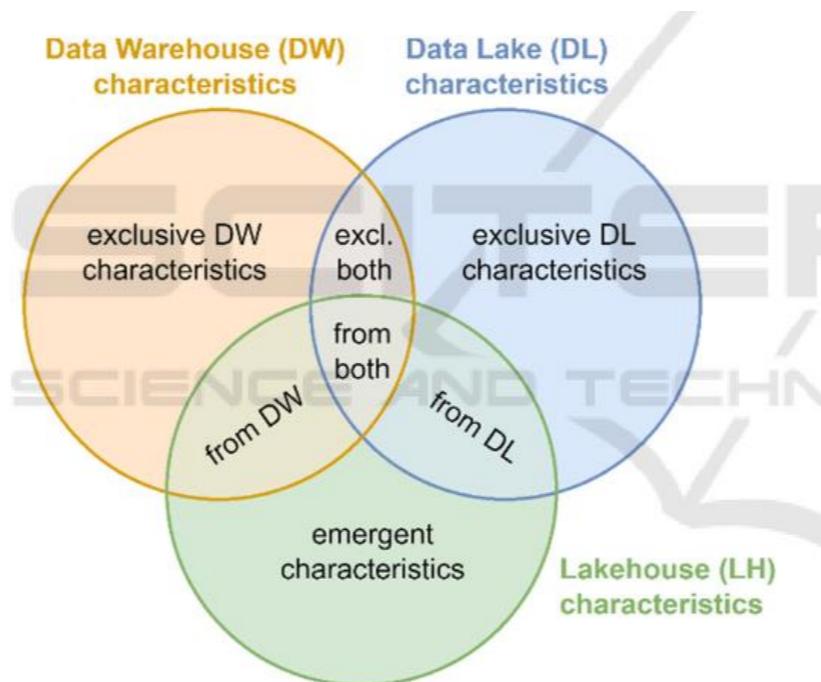


Figure 2: Venn diagram illustrating how the characteristics of lakehouses are composed.

Table 2: Comparison of the analytical lakehouse workloads.

Characteristics	Reporting/OLAP	DM/ML	Streaming
Analytics types:	Descript., diagnostic	Diagnostic, predictive, prescriptive	Descriptive, diagnostic, predictive
Users:	Business users, data analysts	Data scientists	Operators, analysts
Data access:	Via query language	Direct access on storage	Direct acc. on stream storage
Timing:	Batch	Batch	Near-real-time
Data types:	Structured	All types	Structured, semi-struct.
User concurrency:	High	Low	Low

Source:  
<https://www.scitepress.org/Papers/2023/118405/118405.pdf>

Table 3: Overview about the identified lakehouse requirements and the workloads from which they were mainly derived.

Requirement		Influencing workloads		
#	Name	Reporting/OLAP	DM/ML	Streaming
<b>R1</b>	Same storage type and data format	●	●	●
<b>R2</b>	CRUD for all types of data	●	●	●
<b>R3</b>	Relational data collections	●	○	○
<b>R4</b>	Query language	●	○	○
<b>R5</b>	Consistency guarantees	●	○	○
<b>R6</b>	Isolation and atomicity	●	○	●
<b>R7</b>	Direct read access	○	●	○
<b>R8</b>	Unified batch and stream processing	○	○	●

● strong influence      ○ medium influence      ○ no influence

<sup>8</sup> <https://aws.amazon.com/s3/>

<sup>9</sup> <https://azure.microsoft.com/products/storage/blobs/>

Table 4: Evaluation results for six popular data management tools. The numbered columns indicate which requirements are satisfied by each tool, while the last column concludes whether they enable to build lakehouses.

Tool	Version	R1	R2	R3	R4	R5	R6	R7	R8	Lakehouses?
Delta Lake	2.1.0	✓	✓	✓	✓	✓	✓	✓	✓	✓
Apache Hudi	0.12.1	✓	✓	✓	✓	✓	✓	✓	✓	✓
Apache Iceberg	1.0.0	✓	✓	✓	✓	✓	✓	✓	✓	✓
Snowflake with internal tables	6.31.1	✓	✓	✓	✓	✓	✓	✗	✗	✗
Snowflake with external tables	6.31.1	✗	✗	✓	✓	✗	✗	✗	✗	✗
Dremio	23.0.1	✗	✓	✓	✓	✗	✗	✗	✗	✗
Trino	394	✗	✓	✓	✓	✗	✗	✗	✗	✗

<sup>10</sup> <https://flink.apache.org>

<sup>11</sup> <https://snowflake.com>

<sup>12</sup> <https://kafka.apache.org>

### Ale warto skonfrontować cd:

- <https://delta.io/blog/delta-lake-vs-parquet-comparison/>
- Schneider, Jan, et al. "Assessing the Lakehouse: Analysis, Requirements and Definition." *ICEIS* (1). 2023.
- Hassan, I. I. "STORAGE STRUCTURES IN THE ERA OF BIG DATA: FROM DATA WAREHOUSE TO LAKEHOUSE." *Journal of Theoretical and Applied Information Technology* 102.6 (2024).
- Tannier, Xavier, et al. "Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse." *Methods of Information in Medicine* (2024).
- Thomas, Robert, and Julia Anderson. "The Impact of Master Data Management on Business Intelligence and Analysis." (2024).
- Vassiliadis, Panos, et al. "A Conceptual Model for Data Storytelling Highlights in Business Intelligence Environments." *arXiv preprint arXiv:2403.00981* (2024).
- Olaoye, Favour, and Kaledio Potter. *Business Intelligence (BI) and Analytics Software: Empowering Data-Driven Decision-Making*. No. 12550. EasyChair, 2024.
- Juli, Mr. "Enhancing Business Intelligence: Harnessing Text Analytics, AI, and ERP for Data-Driven Decision Making." (2024).
- George, Amrita, Kurt Schmitz, and Veda C. Storey. "A framework for building mature business intelligence and analytics in organizations." *Journal of Database Management (JDM)* 31.3 (2020): 14-39.
- Weichbroth, Paweł, Jozef Zurada, and Celina Olszak. "Exploring the Benefits, Challenges, and Opportunities of Collaborative Business Intelligence." (2024).
- Pedersen, Asbjørn Malte, and Claus Bossen. "Cultivating Data Practices Across Boundaries: How Organizations Become Data-driven." *Computer Supported Cooperative Work (CSCW)* (2024): 1-45.
- Sobkhiz, Soroush, and Tamer El-Diraby. "Natural Language Processing for Building Maintenance: From Deep Learning to Business Intelligence." Available at SSRN 4783740.
- **I wiele wiele innych ...**

# Ćwiczenie 6 – Budowanie tabeli wymiarowej z bazy danych OLTP

Cel Zadania:

Celem tego zadania jest przekształcenie wybranych danych z bazy danych OLTP na schemat wymiarowy, odpowiedni dla systemów OLAP, a następnie zaimplementowanie tego schematu w systemie zarządzania bazą danych (DBMS). Czyli przenieść dane z operacyjnej bazy danych, gdzie nacisk kładziony powinien położony zostać na szybkość i efektywność transakcji, do bazy danych analitycznej, gdzie priorytetem jest szybkość i elastyczność zapytań analitycznych.

Kroki do wykonania:

1. Wybór Zestawu Danych OLTP: Proszę zacząć stworzenia bazy danych OLTP (można stworzyć indywidualnie lub użyć przykładową bazę danych (np. Northwind, AdventureWorks, Kaggle, UCI Machine Learning Repository itd.). Baza danych powinna zawierać przynajmniej tabelę dotyczącą np. klientów, zamówień i produktów (lub inny proces biznesowy).
2. Analiza i Projektowanie Schematu Wymiarowego: Następnym krokiem jest analiza wybranych danych i zaprojektowanie schematu wymiarowego. Proces ten powinien obejmować identyfikację tabel wymiarowych (np. wymiar czasu, klientów, produktów) i tabeli faktów (np. sprzedaży, zamówienia). Opracowanie modeli danych i diagramów schematu wymiarowego, uwzględniając klucze obce, atrybuty wymiarów i miary w tabelach faktów.
3. Eksport i Transformacja Danych: Po zaprojektowaniu schematu wymiarowego, proszę wyeksportować dane z wybranej bazy danych OLTP i przeprowadzić niezbędne transformacje danych, aby dopasować je do nowego schematu. Może to obejmować operacje takie jak czyszczenie danych, agregacja, oraz tworzenie nowych atrybutów wymiarowych.
4. Implementacja Schematu Wymiarowego: Używając wybranego systemu zarządzania bazą danych proszę zaimplementować zaprojektowany schemat wymiarowy, tworząc odpowiednie tabeli wymiarowe i faktowe.
5. Załadowanie Danych i Weryfikacja: Po utworzeniu schematu wymiarowego, proszę zmigrować przekształcone dane do nowych tabel i przeprowadzić weryfikację, aby upewnić się, że dane zostały poprawnie przekształcone i załadowane.
6. Analiza Danych: Proszę zaprezentować wykazanie kilku zapytań SQL na nowo utworzonym schemacie wymiarowym, aby zademonstrować, jak można wykorzystać hurtownię do uzyskiwania wglądów biznesowych. Przykładowe zapytania mogą obejmować analizę trendów sprzedaży w czasie, porównanie wyników sprzedaży między różnymi regionami, czy analizę zachowań zakupowych klientów.
7. Przygotowanie raportu/sprawozdania podsumowującego realizację zadania (opisy, screeny, skrypty) i przesłanie go Prowadzącemu **do 06.05 br.**

# Ćwiczenie 7 – Denormalizacja danych przy użyciu Power Query

## Cel Zadania

Celem tego zadania jest użycie Power Query do denormalizacji danych pochodzących z normalizowanej bazy danych, w tym łączenie tabel, agregowanie danych i przekształcanie wynikowego zestawu danych do formy j dla analizy danych i raportowania.

## Kroki do wykonania:

1. Przygotowanie Danych: zestaw normalizowanych tabel (np..w formacie Excel lub CSV) które zawierają informacje o klientach, produktach i zamówieniach. Możesz użyć danych dostarczonych wcześniej w zadaniu OLTP do OLAP jako punkt wyjścia lub przygotować inny zestaw.
2. Denormalizacja za pomocą Power Query:
  - Importowanie danych: Zainportować każdą tabelę do Power Query.
  - Łączenie tabel: Następnie, używając opcji 'Merge Queries' (łączenie zapytań), proszę połączyć te tabele w jedną denormalizowaną tabelę. Na przykład, można połączyć tabelę Zamówienia z tabelą Klienci na podstawie KlientID oraz tabelę Produkty na podstawie ProduktID
  - Transformacja danych: proszę zastosować odpowiednie transformacje w Power Query, aby wyświetlać szczegóły klientów i produktów bezpośrednio w tabeli zamówień.
3. Agregacja Danych: Używając opcji 'Group By' w Power Query, proszę stworzyć agregacje, takie jak całkowita sprzedaż po produkcie, liczba zamówień na klienta itd.
4. Ładowanie Wynikowych Danych: po zakończeniu transformacji, proszę załadować przekształcone dane z powrotem do arkusza Excela, tworząc tabelę, która będzie służyła jako źródło danych dla raportów i analiz.
5. Analiza i Raportowanie - Na podstawie załadowanych danych proszę utworzyć proste raporty lub dashboardy w Excelu, używając tabel przestawnych lub narzędzi do wizualizacji danych (lub użyć Power BI, lub innego narzędzia)
6. Przygotowanie raportu/sprawozdania podsumowującego realizację zadania (opisy, screeny, skrypty) i przesłanie go Prowadzącemu **do 06.05 br.**

# Ćwiczenie 8 – Operacje w datalakehouse

## Cel:

Celem tego ćwiczenia jest zapoznanie się z podstawowymi operacjami w Data Lakehouse np. przy użyciu narzędzia <https://community.cloud.databricks.com/> (lub innego, wybór uzasadnij). Przygotuj raport/sprawozdania podsumowujące realizację zadania (opisy, screeny, skrypty) i przesłanie go Prowadzącemu do 13.05 br.

Założymy, że firma prowadzi sklep internetowy i gromadzi dane o transakcjach, produktach i klientach. Dane te są przechowywane w postaci plików CSV w Data Lakehouse. Przygotuj krokowy raport wykonalności data pipeline dla tej organizacji, w celu analizy danych sprzedażowych, w tym m.in.:

- Łączenie się z Data Lakehouse: używanie notebooka Databricks, aby połączyć się z Data Lakehouse.
- Ładowanie dane: pliki CSV z danymi o transakcjach, produktach i klientach do Data Lakehouse.
- Oczyszczanie i przetwarzanie danych: Oczyść dane, usuwając duplikaty i błędy. Przetwórz dane, aby przygotować je do analizy.
- Tworzenie tabel: Utwórz tabele dla transakcji, produktów i klientów w Data Lakehouse.
- Analizowanie dane: Wykonaj zapytania SQL, aby przeanalizować dane i uzyskać wgląd w sprzedaż, produkty i klientów.
- Wizualizowanie danych: Użyj narzędzi wizualizacyjnych do stworzenia wykresów i diagramów przedstawiających dane.
- Tworzenie modelu uczenia maszynowego, aby przewidzieć przyszłe zakupy klientów.
- Automatyzowanie procesu ładowania i przetwarzania danych.
- Weryfikacja możliwości udostępnienia danych innym użytkownikom.

## Przydatne zasoby:

- <https://docs.databricks.com/en/index.html>
- <https://community.cloud.databricks.com/>
- [https://m.youtube.com/watch?v=4y6g0Vyc\\_PY](https://m.youtube.com/watch?v=4y6g0Vyc_PY)
- <https://docs.databricks.com/en/getting-started/etl-quick-start.html>
- <https://docs.databricks.com/en/getting-started/data-pipeline-get-started.html>

# Ćwiczenie 9 – Dax

## Opis zadania:

Twoim zadaniem jest przeprowadzenie analizy danych dotyczących sprzedaży w fikcyjnej firmie handlowej. Dane wejściowe (wzór w załączniku) jakimi dysponujesz w organizacji, mogą wymagać dostosowania. Aktualnie magazynowane są w Excel. Zadanie polega na stworzeniu raportu zawierającego różne wskaźniki i analizy dotyczące sprzedaży oraz przedstawienie sposobu realizacji w formie sprawozdania.

## Kroki do wykonania:

### 1. Import danych:

- Zimportuj dane z pliku Excela/lub dolinkuj DB do Power BI.
- Upewnij się, że dane są poprawnie zinterpretowane przez program i zrozumiane jako odpowiednie typy danych.

### 2. Czyszczenie danych:

- Zidentyfikuj i usuń lub zastąp brakujące wartości w danych, jeśli istnieją.
- Sprawdź, czy dane są poprawne i spójne. W razie potrzeby dokonaj korekty.

### 3. Tworzenie kalkulacji w języku DAX:

- Stwórz kolumny obliczeniowe, które będą zawierać:
  - Przychód ze sprzedaży (Sales Revenue)
  - Sumy sprzedaży dla poszczególnych produktów
  - Średnie miesięczne sprzedaży.
  - Średnie roczne sprzedaży.
  - Procentowy udział każdego produktu w całkowitej sprzedaży.
  - Średnia wartość zamówienia (AOV) - średnia kwota pieniędzy wydana przez klienta na jedno zamówienie
  - Wartość życia klienta (CLTV) - łączna kwota pieniędzy, jaką klient wyda w firmie w ciągu całego okresu współpracy (definiowanego i ad hoc y-m-d) - pomiar rentowności klientów.
  - Inne wskaźniki, które uznasz za istotne dla analizy sprzedaży, lub będą niezbędne do pełnego wnioskowania.

### 4. Wizualizacja danych:

- Stwórz różne wizualizacje, takie jak wykresy słupkowe, kołowe, liniowe itp., aby przedstawić wyniki analizy sprzedaży (ww KPI).
- Dodaj filtry, które pozwolą użytkownikowi interaktywnie eksplorować dane, na przykład filtrowanie według kategorii produktów, regionów sprzedaży itp.

### 5. Analiza trendów:

- Wykorzystaj język M do tworzenia kolumny zawierającej daty w formacie, który umożliwia analizę trendów sprzedaży w kolejnych miesiącach lub latach (łatwość wyodrębnienia roku, miesięcy itd., użycie funkcji group by do obliczenia sumy wartości sprzedaży dla każdego miesiąca i roku itp.).
- Przeprowadź analizę trendów i zidentyfikuj ewentualne wzorce w sprzedaży.

### 6. Ostateczny raport:

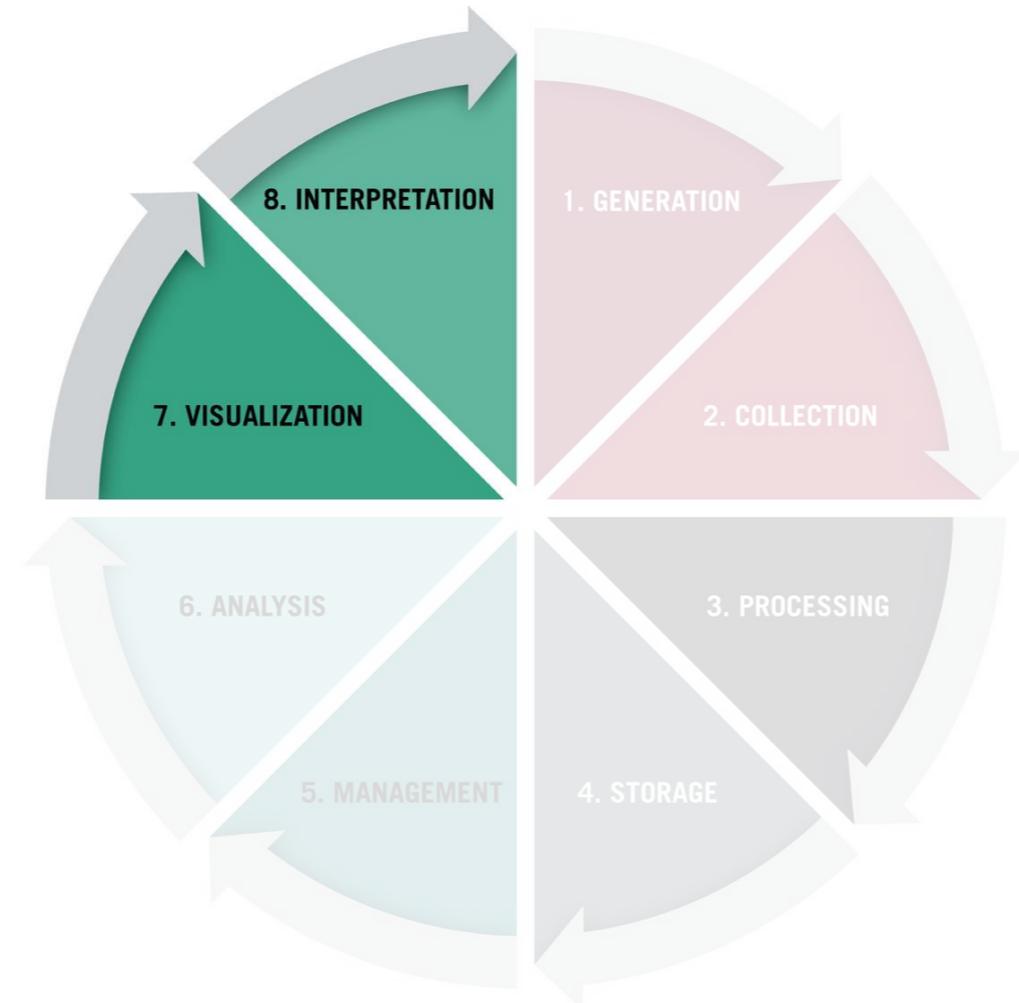
- Przygotuj raport zawierający analizy, wykresy i wnioski związane z analizą danych sprzedażowych.
- Upewnij się, że raport jest czytelny i łatwy do zrozumienia dla osób, które nie są specjalistami od analizy danych.
- Udostępnij i prześlij rezultaty Prowadzącemu;

**Zasoby:** Dokumentacja Power BI, języka DAX i M, np. :

- <https://learn.microsoft.com/pl-pl/dax/>,  
<https://learn.microsoft.com/en-us/powerquery-m/> etc.

Termin wykonania: 08.05.2024 – 15.05.2024

Dokumenty	General	Cw_Dax (08.05.2024)	Zmodyfikowane	Wyszczególnić	Dodaj do katalogu



Tworzenie wizualizacji BI to kluczowy proces, który umożliwia organizacjom transformację danych w wiedzę, wspierającą podejmowanie decyzji (interpretację). Wizualizacje - Raporty BI powinny być przygotowywane z uwzględnieniem kilku fundamentalnych założeń:

**Zrozumienie celu i odbiorców:** Każdy raport powinien być zaprojektowany z myślą o konkretnym celu oraz grupie odbiorców. To zrozumienie wpływa na zawartość, poziom szczegółowości, a także sposób prezentacji danych. Raporty dla kierownictwa często są bardziej strategiczne i skondensowane, natomiast raporty operacyjne mogą zawierać więcej szczegółów i być aktualizowane częściej

**Dostępność i integracja danych:** Raporty BI powinny być oparte na danych, które są dostępne i mogą być łatwo zintegrowane z różnych źródeł. Ważne jest, aby zapewnić, że dane są aktualne, dokładne i spójne

**Interaktywność i użyteczność:** Nowoczesne narzędzia BI oferują możliwość tworzenia interaktywnych raportów, które umożliwiają użytkownikom samodzielna analizę danych poprzez filtrowanie, sortowanie czy eksplorację szczegółów. Interaktywność zwiększa użyteczność raportów, umożliwiając odbiorcom głębsze zrozumienie prezentowanych informacji

**Wizualizacja danych:** Dobre wizualizacje są kluczowe dla skuteczności raportów BI. Wykresy, grafy, mapy i dashboardy powinny być stosowane, aby uprościć zrozumienie złożonych zestawień danych i uwydatnić kluczowe wskaźniki (KPIs). Wizualizacja powinna być przejrzysta i dostosowana do rodzaju danych oraz celu raportu

**Automatyzacja i skalowalność:** Raporty BI powinny być łatwe w utrzymaniu i aktualizacji. Automatyzacja procesu zgromadzenia danych, ich przetwarzania oraz dystrybucji raportu jest kluczowa, zwłaszcza w dynamicznie zmieniających się środowiskach. Skalowalność rozwiązania pozwala na rozbudowę i modyfikację raportów w miarę ewolucji potrzeb organizacji

**Regularne przeglądy i aktualizacje:** Świat biznesu szybko się zmienia, więc raporty BI również powinny być regularnie przeglądane i aktualizowane, aby zapewnić, że nadal odpowiadają na aktualne potrzeby i wyzwania biznesowe.

Zrozumienie celu i odbiorców jest kluczowym aspektem w tworzeniu efektywnych raportów i narzędzi analitycznych w Business Intelligence (BI). To zrozumienie pozwala na dopasowanie raportu do specyficznych potrzeb i oczekiwania danej grupy użytkowników, co znacząco wpływa na jego skuteczność.

## Zasady

- **Dokładna identyfikacja potrzeb odbiorców:** Zrozumienie, co odbiorcy chcą osiągnąć za pomocą raportu, jest kluczowe. Czy potrzebują oni informacji do podejmowania strategicznych decyzji? A może chcą monitorować codzienne operacje? Rozumienie tych potrzeb pozwala na skonstruowanie treści i formatu raportu, które będą najbardziej pomocne.
- **Personalizacja komunikacji:** Dostosowanie języka, terminologii i poziomu szczegółowości do wiedzy i doświadczenia odbiorców. Dla kierownictwa wyższego szczebla zazwyczaj preferowane są skondensowane, strategiczne przekazy, podczas gdy specjalisci techniczni mogą potrzebować bardziej szczegółowych danych.
- **Zapewnienie interaktywności i elastyczności:** Dostarczenie użytkownikom narzędzi umożliwiających eksplorację danych na własną rękę, co pozwala im na głębsze zrozumienie i lepsze dostosowanie informacji do własnych potrzeb.

**1. Stakeholderzy:** Wszystkie osoby lub grupy, które mają interes w raporcie. Rozumienie ich roli w organizacji i tego, jak będą używać raportów, jest niezbędne.

**2. User Persona:** Fikcyjne profile użytkowników reprezentujące różne segmenty odbiorców. Persona pomaga zrozumieć motywacje, potrzeby i ograniczenia poszczególnych użytkowników, co ułatwia tworzenie bardziej celowanych i efektywnych raportów.

**3. User Experience (UX):** Projektowanie doświadczeń użytkownika w kontekście interakcji z rapportami BI, tak aby były one intuicyjne, użyteczne i estetycznie przyjemne.

## Metody

- **Ankiety i wywiady z użytkownikami:** Bezpośrednie zbieranie informacji od przyszłych użytkowników raportów, które pomagają zrozumieć ich oczekiwania, doświadczenia i preferencje.
- **Workshops (Warsztaty):** Sesje robocze z użytkownikami końcowymi, które pomagają w głębszym zrozumieniu ich codziennych zadań i wyzwań. Warsztaty mogą również służyć do testowania prototypów raportów.
- **Obserwacje i analiza zachowań:** Monitoring, jak użytkownicy korzystają z obecnych systemów BI i rapportów, co może dostarczyć wskazówek, jak ulepszyć nowe rozwiązania.

Zrozumienie celu i odbiorców w kontekście tworzenia rapportów BI wymaga szczegółowego badania i analizy potrzeb i oczekiwaniń wszystkich zainteresowanych stron. Podejście to nie tylko zwiększa użyteczność końcowego produktu, ale także przyczynia się do bardziej efektywnego wykorzystania zasobów organizacji poprzez dostarczenie wartościowych i skrojonych na miarę informacji decyzyjnych.

- **Stakeholderzy:** Wszystkie osoby lub grupy, które mają interes w raporcie. Rozumienie ich roli w organizacji i tego, jak będą używać raportów, jest niezbędne.
  - Top Management – strategiczne, wysokopoziomowe informacje.
  - Management – big picture, ale bardziej detaличny niż Top Management.
  - Zespoły – zwykle najbardziej detaличne informacje operacyjne.
- **User Persona:** Fikcyjne profile użytkowników reprezentujące różne segmenty odbiorców. Persona pomaga zrozumieć motywacje, potrzeby i ograniczenia poszczególnych użytkowników, co ułatwia tworzenie bardziej celowanych i efektywnych raportów. -> UX

## Metody

- **Ankiety i wywiady z użytkownikami:** Bezpośrednie zbieranie informacji od przyszłych użytkowników raportów, które pomagają zrozumieć ich oczekiwania, doświadczenia i preferencje.
- **Workshops (Warsztaty):** Sesje robocze z użytkownikami końcowymi, które pomagają w głębszym zrozumieniu ich codziennych zadań i wyzwań. Warsztaty mogą również służyć do testowania prototypów raportów.
- **Obserwacje i analiza zachowań:** Monitoring, jak użytkownicy korzystają z obecnych systemów BI i raportów, co może dostarczyć wskazówek, jak ulepszyć nowe rozwiązania.

Zrozumienie celu i odbiorców w kontekście tworzenia raportów BI wymaga szczegółowego badania i analizy potrzeb i oczekiwanią wszystkich zainteresowanych stron. Podejście to nie tylko zwiększa użyteczność końcowego produktu, ale także przyczynia się do bardziej efektywnego wykorzystania zasobów organizacji poprzez dostarczenie wartościowych i skrojonych na miarę informacji decyzyjnych.

# UX tutaj też ma znaczenie ...

Istnieje wiele metod UX, które można stosować na różnych etapach procesu projektowania.

## Badania:

- **Wywiady:**
  - Wywiady pogłębione (IDI)
  - Wywiady grupowe (FGI)
  - Wywiady z ekspertami
  - Testy użyteczności
- **Badania ankietowe:**
  - Ankiety online
  - Ankiety papierowe
- **Badania obserwacyjne:**
  - Obserwacje terenowe
  - Eyetracking
  - Analiza nagrąń sesji
- **Badania etnograficzne:**
  - Badania kontekstowe
  - Badania diary
  - Mapping podróży użytkownika

## Analiza:

- **Analiza heurystyczna:** Ocena projektu na podstawie uznanych zasad użyteczności.
- **Analiza konkurencji:** Badanie produktów i usług konkurentów.
- **Analiza danych:** Analiza danych użytkowania w celu zidentyfikowania wzorców i możliwości usprawnień.
- **Sortowanie kart:** Użytkownicy sortują karty z etykietami w celu określenia struktury informacji.

## Projektowanie:

- **Szkicowanie:** Szybkie tworzenie pomysłów na papierze.
- **Storyboarding:** Tworzenie wizualnej historii interakcji użytkownika z produktem.
- **Prototypowanie:** Tworzenie interaktywnych modeli produktu w celu testowania i zbierania opinii (low fidelity/full fidelity mockups)
- **Tworzenie map mentalnych:** Wizualizowanie pomysłów i ich powiązań.
- **Warsztaty projektowe:** Wspólne opracowywanie pomysłów z zespołem i użytkownikami.

## Testowanie:

- **Testy użyteczności:** Obserwowanie użytkowników podczas wykonywania zadań w celu zidentyfikowania problemów.
- **Testy A/B:** Porównywanie dwóch wersji projektu w celu określenia, która z nich działa lepiej.
- **Testy z udziałem użytkowników:** Testowanie produktu z rzeczywistymi użytkownikami w celu uzyskania opinii.
- **Testy odbioru:** Testowanie produktu przez użytkowników docelowych przed jego wydaniem.

## Ewaluacja:

- **Ocena heurystyczna:** Ocena projektu na podstawie uznanych zasad użyteczności.
- **Analiza danych:** Analiza danych użytkowania w celu zidentyfikowania wzorców i możliwości usprawnień.
- **Testy użyteczności:** Obserwowanie użytkowników podczas wykonywania zadań w celu zidentyfikowania problemów.
- **Ankiety satysfakcji:** Zbieranie opinii użytkowników na temat produktu.

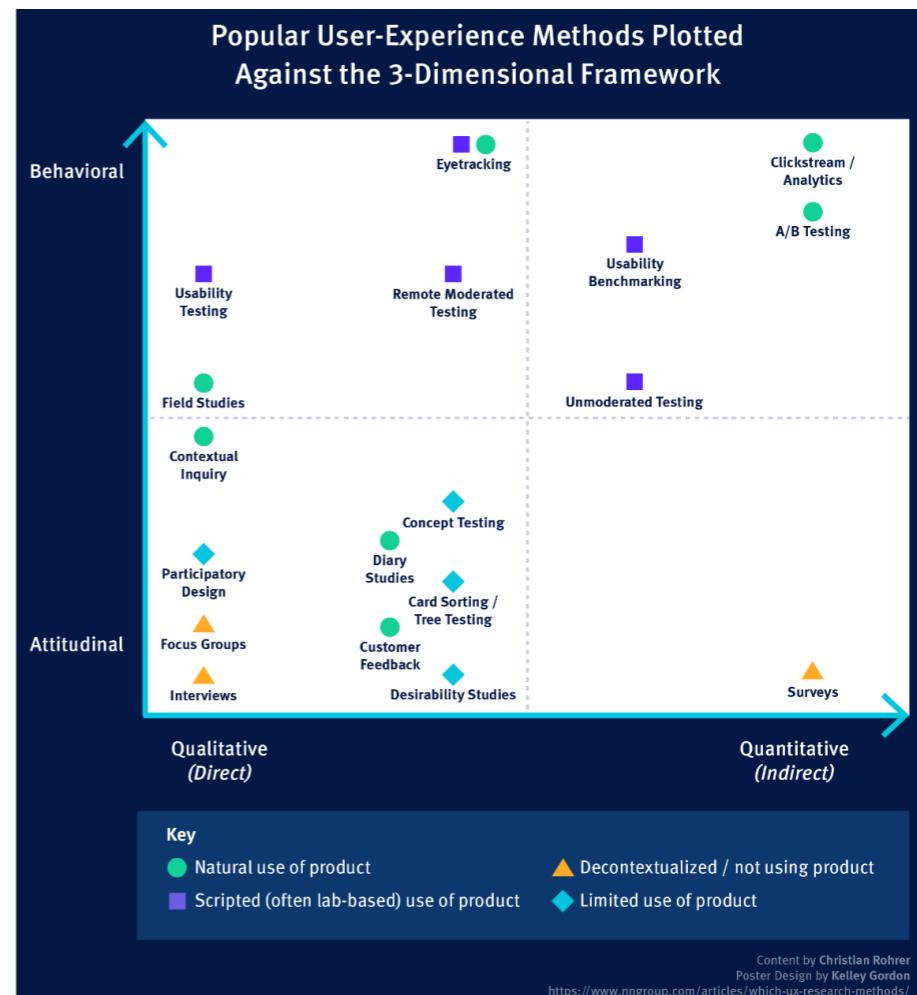
## Zrozumienie celu i odbiorców

### 20 UX Methods in Brief

Key	
	▲ Decontextualized / not using product
	◆ Limited use of product
■ <b>Usability Testing</b>	Participants are brought into a lab, one-on-one with a researcher, and given a set of scenarios that lead to tasks and usage of specific interest within a product or service.
● <b>Field Studies</b>	Researchers study participants in their own environment (work or home), where they would most likely encounter the product or service being used in the most realistic or natural environment.
● <b>Contextual Inquiry</b>	Researchers and participants collaborate together in the participants own environment to inquire about and observe the nature of the tasks and work at hand.
◆ <b>Participatory Design</b>	Participants are given design elements or creative materials in order to construct their ideal experience in a concrete way that expresses what matters to them most and why.
▲ <b>Focus Groups</b>	Groups of 3–12 participants are led through a discussion about a set of topics, giving verbal and written feedback through discussion and exercises.
▲ <b>Interviews</b>	A researcher meets with participants one-on-one to discuss in depth what the participant thinks about the topic in question.
◆ <b>Concept Testing</b>	A researcher shares an approximation of a product or service that captures the key essence (the value proposition) of a new concept or product in order to determine if it meets the needs of the target audience. It can be done one-on-one or with larger numbers of participants, and either in person or online.
● <b>Diary Studies</b>	Participants are given a mechanism (diary or camera) to record and describe aspects of their lives that are relevant to a product or service or simply core to the target audience. Diary studies are typically longitudinal and can be done only for data that is easily recorded by participants.
● <b>Customer Feedback</b>	Open-ended and/or close-ended information provided by a self-selected sample of users, often through a feedback link, button, form, or email.
◆ <b>Desirability Studies</b>	Participants are offered different visual-design alternatives and are expected to associate each alternative with a set of attributes selected from a closed list. These studies can be both qualitative and quantitative.
▲ <b>Card Sorting</b>	A quantitative or qualitative method that asks users to organize items into groups and assign categories to each group. This method helps create or refine the information architecture of a site by exposing users' mental models.

<https://www.nngroup.com/articles/guide-ux-research-methods/>

## UX tutaj też ma znaczenie ...



## Zrozumienie celu i odbiorców

### NN/g Nielsen Norman Group

World Leaders in Research-Based User Experience

Log in

Search

Home Articles Training & UX Certification Consulting Reports & Books About NN/g

#### Topics

Agile  
Artificial Intelligence  
Design Process  
Ecommerce  
Intranets  
Navigation  
Psychology and UX  
Research Methods  
Study Guides  
User Testing  
Web Usability  
Writing for the Web

▷ See all topics  
  
Popular Articles  
10 Usability Heuristics for User Interface Design  
Empathy Mapping: The First Step in Design Thinking  
When to Use Which User-Experience Research Methods  
Service Blueprints: Definition  
Journey Mapping 101  
The Four Dimensions of Tone of Voice  
Between-Subjects vs.

#### Level Up Your Focus Groups

**Summary:** To avoid common issues with focus groups, consider using these strategies to maximize participation and minimize the potential for bias.

5 minute video by Therese Fessenden

Topics: Research Methods

Share this video:  
Twitter  
LinkedIn  
Email



Design Taste vs.  
Technical Skills in the Era of AI

Typography Terms:  
Glossary

UX Writing: Study Guide

7 Tips for Memorable and Easy-to-Understand Imagery

Content Standards in Design Systems

#### Recent Videos

What is UX (Not)?

Is Livestream Selling Right for Your Business?

Cookie Permissions: 6 Design Guidelines

Successful Projects: 7 Steps for Better Collaboration

The 3 Competencies of Journey Management

#### Authors

Bruce Tognazzini

Don Norman

Jakob Nielsen

▷ See all authors

#### Learn More

Subscribe to the weekly newsletter to get notified about future articles.

#### Videos



Always Pilot Test User Research Studies  
3 minute video



Inductively Analyzing Qualitative Data  
3 minute video



What, When, Why: Research Goals, Questions, and Hypotheses  
3 minute video



Informed Consent for UX Research  
5 minute video

#### UX Conference Training Course

Discovery: Building the Right Thing

Measuring UX and ROI

Facilitating UX Workshops

Analytics and User Experience

ResearchOps: Scaling User Research

<https://www.nngroup.com/articles/guide-ux-research-methods/>

Dostępność i integracja danych to kluczowe obszary w kontekście systemów BI, które mają bezpośredni wpływ na skuteczność analiz i raportów.

### Zasady

- **Kompletność danych:** Zapewnienie, że wszystkie niezbędne dane są dostępne dla użytkowników systemu BI, co oznacza integrację danych z różnych źródeł i systemów w jednym, spójnym repozytorium.
- **Aktualność danych:** Dane powinny być regularnie aktualizowane, aby odzwierciedlały najnowsze informacje, co jest kluczowe dla podejmowania trafnych decyzji biznesowych.
- **Spójność danych:** Dane powinny być spójne w całej organizacji, bez względu na to, gdzie są przechowywane lub w jaki sposób są przetwarzane. Konieczne jest stosowanie standardów i procedur, które pomagają unikać duplikacji i rozbieżności.
- **Dostępność danych:** Dane powinny być łatwo dostępne dla uprawnionych użytkowników, ale równocześnie chronione przed dostępem nieautoryzowanym. Zarządzanie dostępem i bezpieczeństwem jest tutaj kluczowe.

- **Data Integration:** Zastosowanie narzędzi i technologii (jak ETL) do łączenia danych z różnych źródeł w jedną spójną bazę, co umożliwia bardziej kompleksowe analizy.
- **Data Governance:** Stosowanie polityk, procedur i standardów, które zarządzają dostępnością, integralnością i bezpieczeństwem danych w organizacji.
- **Data Virtualization:** Technologia, która umożliwia szybki dostęp do danych z różnych źródeł bez potrzeby ich fizycznego przenoszenia do jednej lokalizacji, co znacznie usprawnia integrację.
  - **Data Federation:** Metoda, która pozwala na konsolidację i wirtualne wyświetlanie danych z różnych źródeł, tak jakby pochodziły z jednego źródła, bez fizycznego przenoszenia danych.

Dostępność i integracja danych są fundamentem efektywnego wykorzystania systemów BI, ponieważ tylko dobrze zintegrowane i zarządzane dane mogą dostarczyć wartościowych insightów niezbędnych do podejmowania decyzji biznesowych.

Interaktywność i użyteczność są kluczowymi elementami w projektowaniu narzędzi BI, które mają na celu umożliwienie użytkownikom efektywnego dostępu i interpretacji danych.

## Zasady

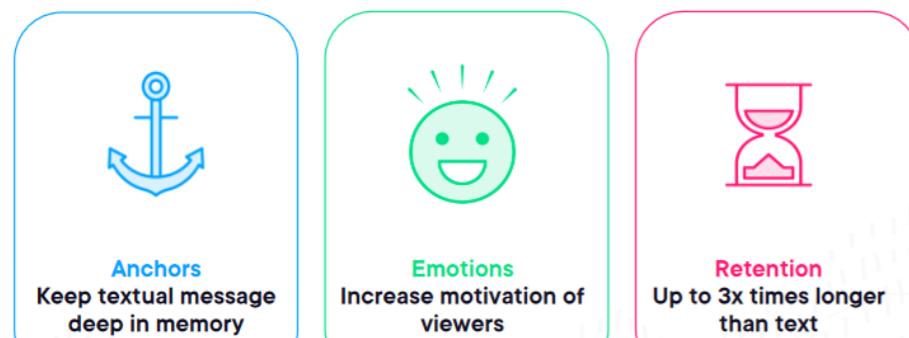
- **Użyteczność (Usability):** Projektowanie rozwiązań BI powinno być zorientowane na użytkownika, co oznacza, że narzędzia powinny być intuicyjne, łatwe w obsłudze i dostosowane do potrzeb użytkownika. Wysoka użyteczność zwiększa zaangażowanie użytkowników i efektywność pracy.
- **Dostosowywanie do kontekstu użytkownika:** Narzędzia BI powinny być dostosowane do specyficznych zadań, które użytkownicy mają wykonać, a interfejs powinien odpowiadać na ich konkretne potrzeby analityczne.
- **Responsywność:** Rozwiązania BI powinny być responsywne, tzn. dobrze działać na różnych urządzeniach i platformach, co jest szczególnie ważne w kontekście rosnącego wykorzystania urządzeń mobilnych w biznesie.
- **Interaktywność:** Możliwość dynamicznego wpływania na prezentowane dane poprzez manipulowanie elementami interfejsu, takimi jak filtry, sortowania czy drill-down (zagłębianie się w szczegóły danych)

**Wizualizacja danych:** Dobre wizualizacje są kluczowe dla skuteczności raportów BI. Wykresy, grafy, mapy i dashboardy powinny być stosowane, aby uprościć zrozumienie złożonych zestawień danych i uwypatnić kluczowe wskaźniki (KPIs). Wizualizacja powinna być przejrzysta i dostosowana do rodzaju danych oraz celu raportu

## Zasady:

- **Jasność i czytelność:** Wizualizacje powinny być łatwe do zrozumienia na pierwszy rzut oka. Unikaj przekomplikowanych grafik, które mogą zaciemniać przekaz.
- **Odpowiedni dobór wizualizacji:** Wybieraj typ wizualizacji odpowiedni do natury danych i celu raportu. Na przykład, wykresy liniowe są dobre do prezentacji trendów, a wykresy kołowe do pokazywania proporcji.
- **Konsystencja:** Używaj spójnej palety kolorów, stylów czcionek i układów w całym raporcie, aby ułatwić odbiór i porównywanie danych między różnymi sekcjami raportu.
- **Minimalizm:** Unikaj zbędnych elementów graficznych, które mogą rozpraszać uwagę od kluczowych informacji. Skup się na danych, eliminując nieistotne ozdobniki.

### Why are Visuals so Powerful?



## Order Design

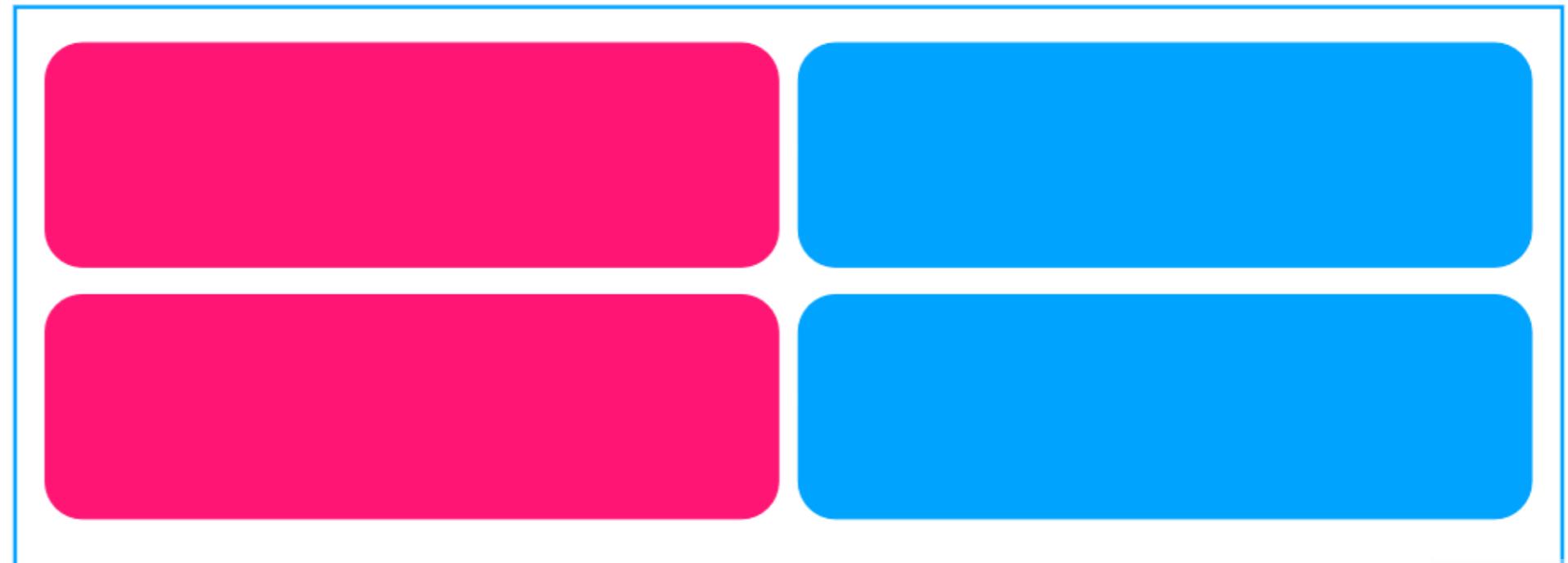


*Gutenberg principle*

Nikola Ilic  
Data Mozart

@DataMozart | www.data-mozart.com

## Balance - Symmetrical

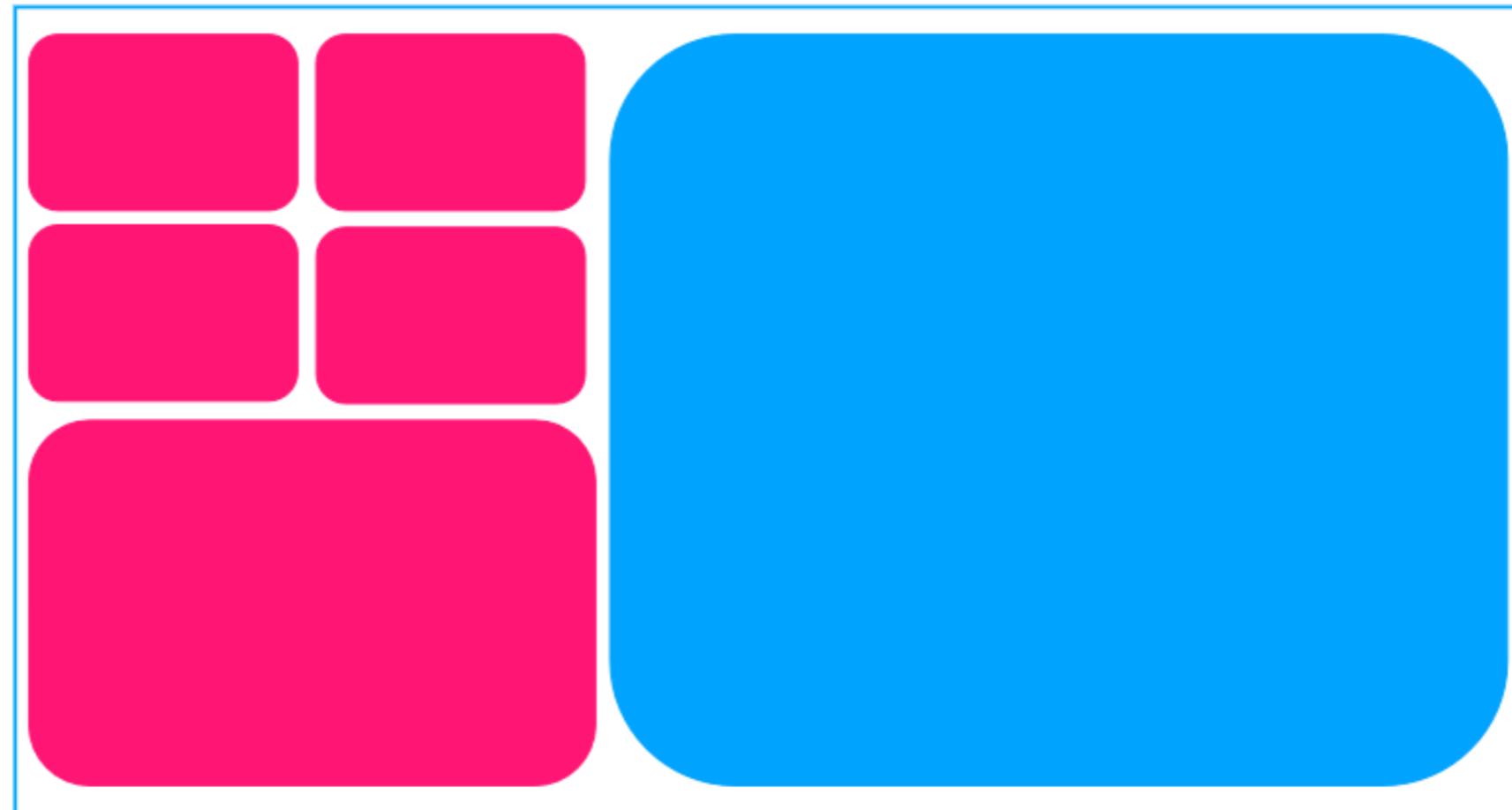


Nikola Ilic

Data Mozart

@DataMozart | www.data-mozart.com

## Balance - Asymmetrical

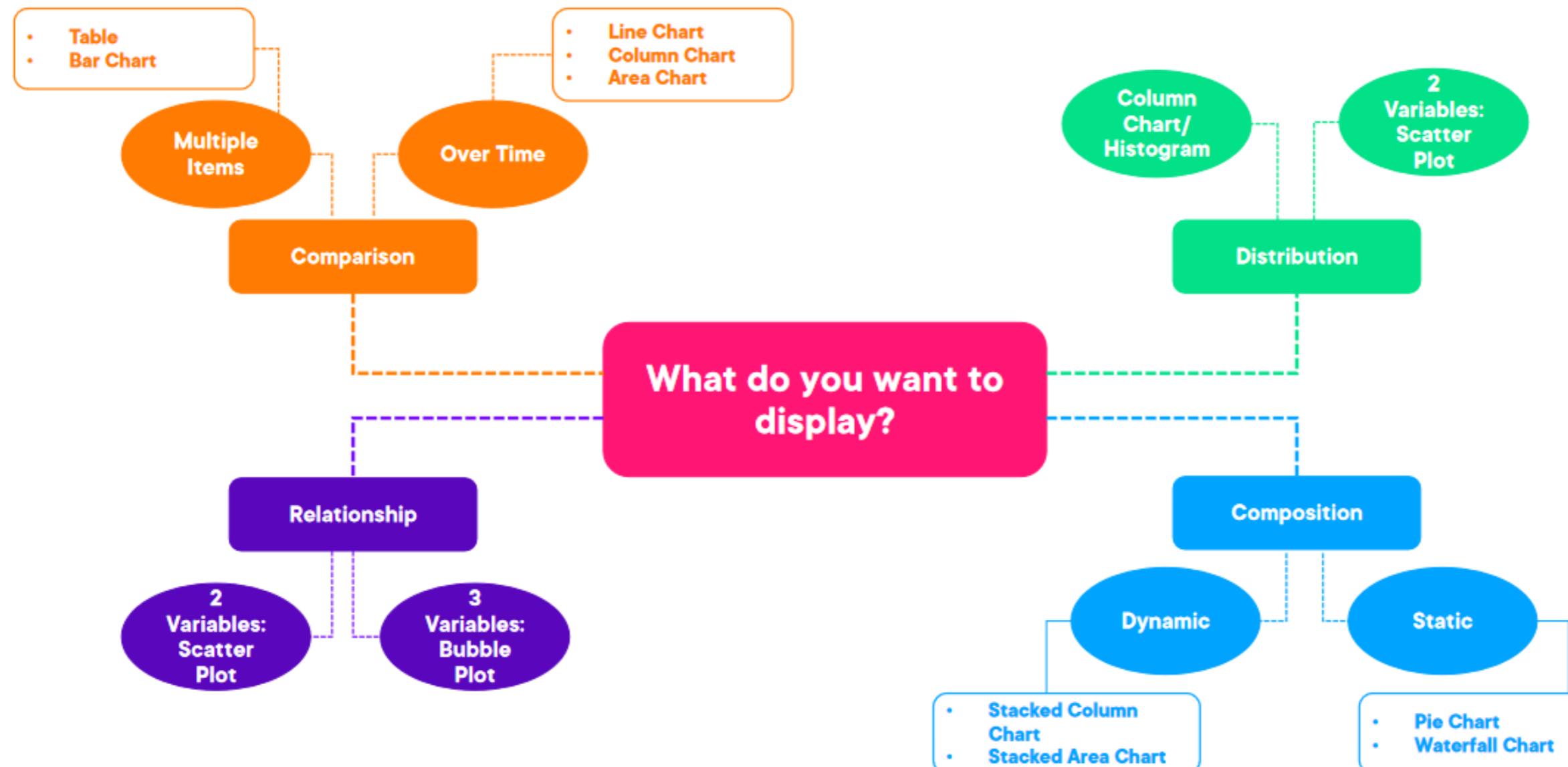


Nikola Ilic  
Data Mozart

@DataMozart | www.data-mozart.com

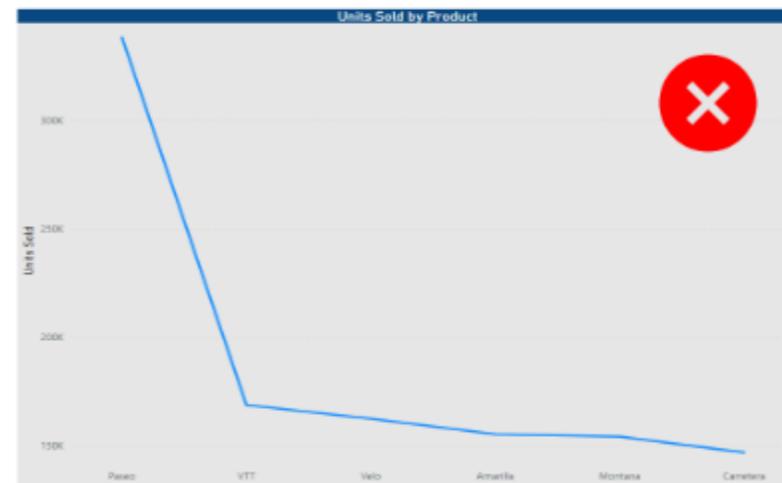
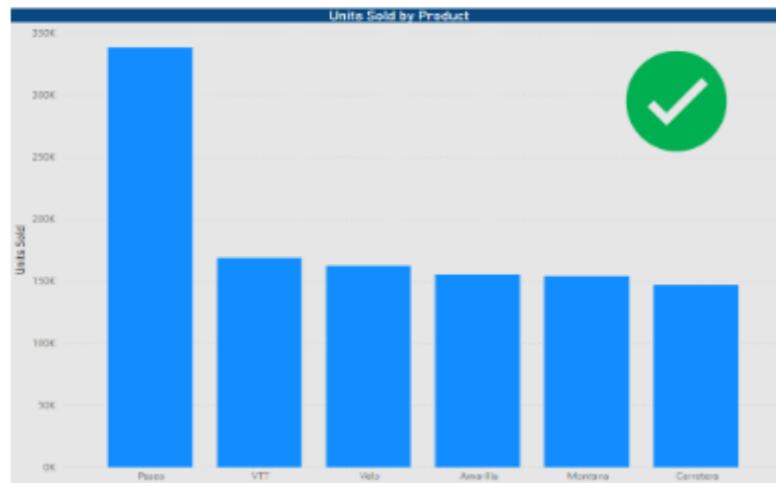
## Contrast and Proximity





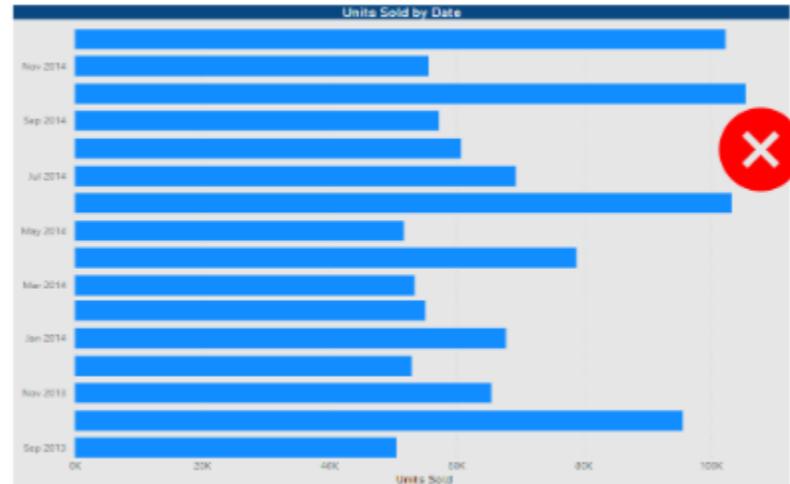
# Displaying Categorical Data

- ✓ Data in groups
- ✓ No logical sequence
- ✓ Easy comparison between data points



# Displaying Changes Over Time

✓ Display values from left to right



Nikola Ilic  
Data Mozart

@DataMozart | www.data-mozart.com

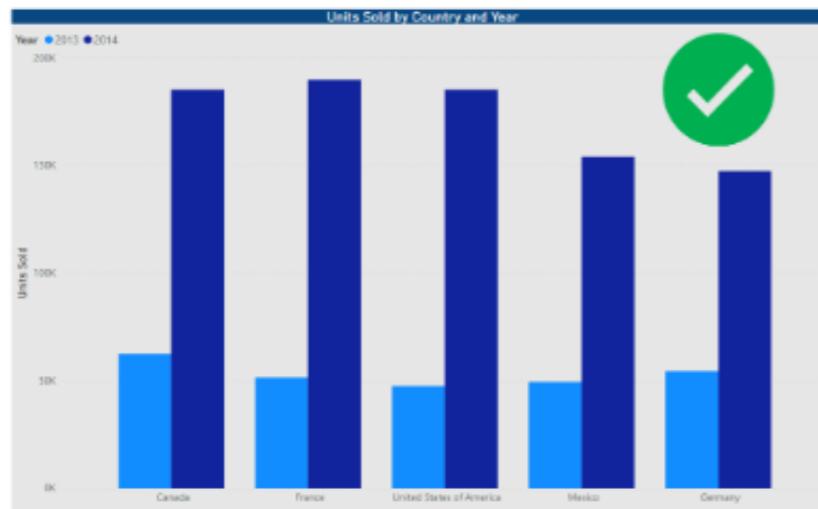
# Data Sorting

- ✓ Alphabetical sorting may be confusing
- ✓ Additional emphasis by using conditional formatting



# Displaying Multi-Dimensional Comparison

✓ Visualize the data by multiple different dimensions (Product & Date)



Nikola Ilic

Data Mozart

@DataMozart | www.data-mozart.com

# Displaying Similar Data Values

✓ Multiple data points with similar values

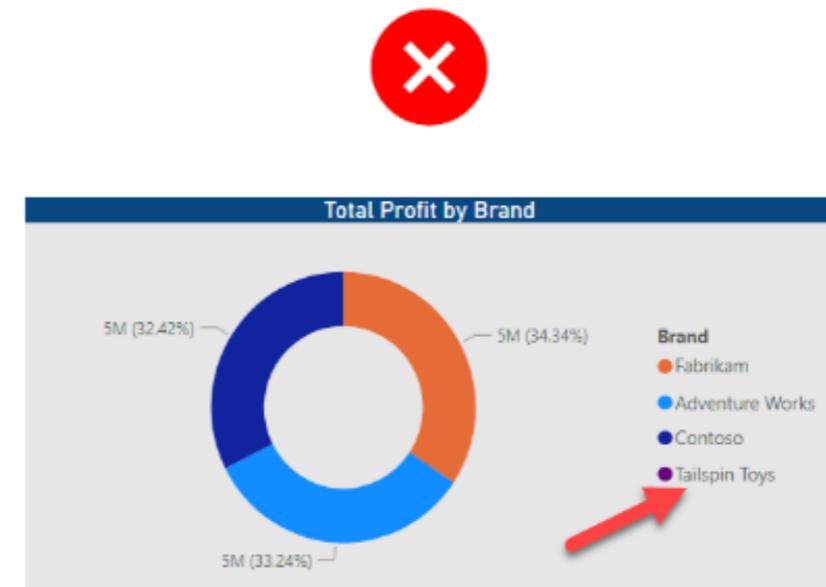
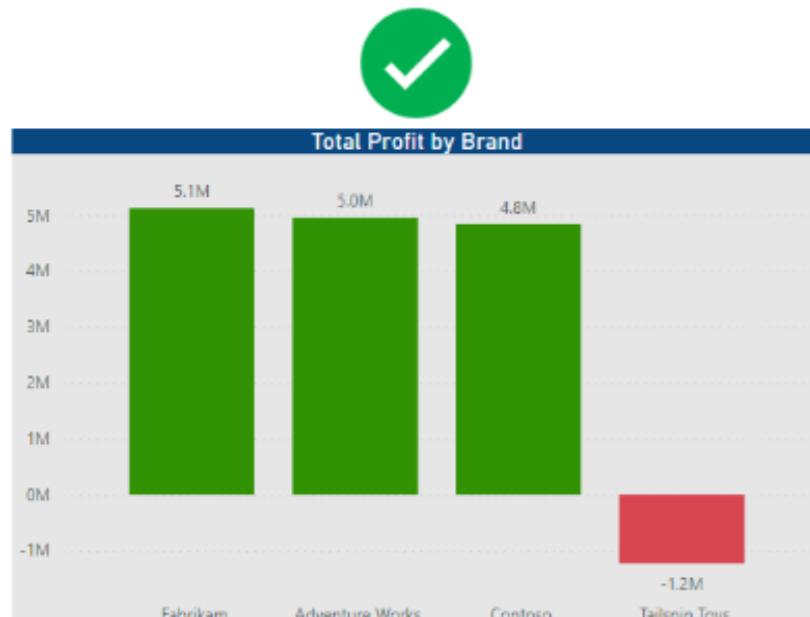


Brand	Total Profit
Fabrikam	5,121,187
Adventure Works	4,958,333
Contoso	4,835,758
<b>Total</b>	<b>14,915,278</b>



# Displaying Negative Values

- ✓ Never use proportional visuals (donuts, pie charts)



# Dostępność dla jak największej liczby użytkowników.



No visual impairments



Green-Blind (Deutanopia)

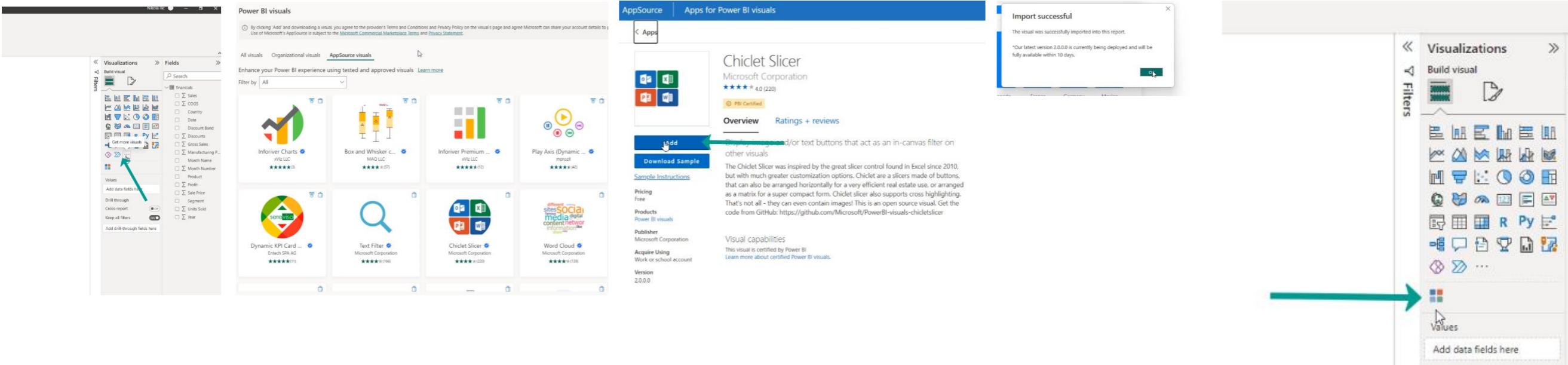
## Accessibility Design Checklist



- Ensure color contrast is at least 4.5 : 1
- Avoid using colors only for transmitting info
- Use clear textual descriptions
- Add Alt text to all non-decorative visuals
- Ensure that the report works for users with visual deficiencies
- Avoid tooltips for conveying important information

## Wizualizacja danych

# Nie zapominamy o Marketplace (np. Power BI)



The collage illustrates the workflow for adding a Power BI visual:

- Power BI Visuals:** Shows the 'Values' section of the Power BI canvas, with a green arrow pointing to the 'Get more visuals' button.
- AppSource | Apps for Power BI visuals:** Displays the 'AppSource visuals' tab, listing various Power BI visual add-ons like 'Inforiver Charts', 'Box and Whisker c...', 'Play Axis (Dynamic ...)', 'Dynamic KPI Card ...', 'Text Filter', 'Chiclet Slicer', and 'Word Cloud'. The 'Chiclet Slicer' item is highlighted.
- Chiclet Slicer Product Page:** Shows the product details for 'Chiclet Slicer' by Microsoft Corporation, including a 4.0 rating (220 reviews), 'Add' button, and a detailed description of the visual's capabilities.
- Import successful:** A modal window confirming the import of the 'Chiclet Slicer' visual into the report.
- Power BI Canvas:** The final view showing the newly imported 'Chiclet Slicer' visual in the Power BI canvas, with a green arrow pointing to its icon.

## Podstawowe pojęcia – przypomnienie:

- Analiza opisowa - statystyki opisowe, statystki podstawowe – jeszcze nie dają konkluzji,
- Analiza inferencyjna (ang. inferential analysis) - metody inferencyjne w statystyce to techniki wykorzystywane do wnioskowania o populacji na podstawie próby,
- Analiza predykcyjna (ang. predictive analysis) – używamy danych historycznych do predykcji nowych zjawisk,
- Analiza diagnostyczna (ang. diagnostic analysis) – ustalenie przyczyny zjawiska w oparciu o dostępne dane,
- Analiza preskryptywna (ang. prescriptive analysis) - nie tylko przewiduje przyszłe wyniki, ale także proponuje działania, które mogą prowadzić do optymalnych wyników. Jest to najbardziej zaawansowana forma analizy biznesowej, łącząca w sobie elementy analizy diagnostycznej, predykcyjnej i decyzyjnej (np. maksymalizacja zysków i minimalizacja kosztów związanych z zarządzaniem zapasami w sieci sklepów).

Charakterystyki/wzorce danych, Data Insights:

- wzorce, trendy (patterns, trends)
- anomalie, wartości odstające (anomalies, outliers)
- Istotne zależności i korelacje (relationships, correlations)
- Identyfikacja, uwagi:
  - Przez analizę opisową, IQR (**Rozstęp międzykwartylowy , czyli 3-1 kwartyl**), np. do detekcji wartości odstających
  - Uwaga: korelacja, [korelacja to nie przyczynowość](#):

*Jeśli czynnik A (np. wykształcenie) i czynnik B (np. zarobki) korelują ze sobą, to powinno się tworzyć przynajmniej kilka hipotez na temat ewentualnego związku przyczynowego między nimi:*

**1.Czynnik A wpływa na czynnik B.** Tu: wykryto związek między zarobkami a wykształceniem, bo wyższe wykształcenie powoduje że dana osoba więcej zarabia.

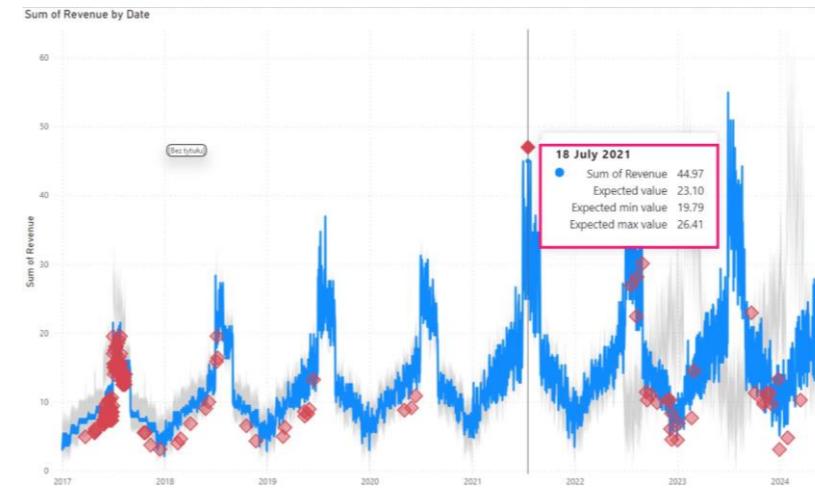
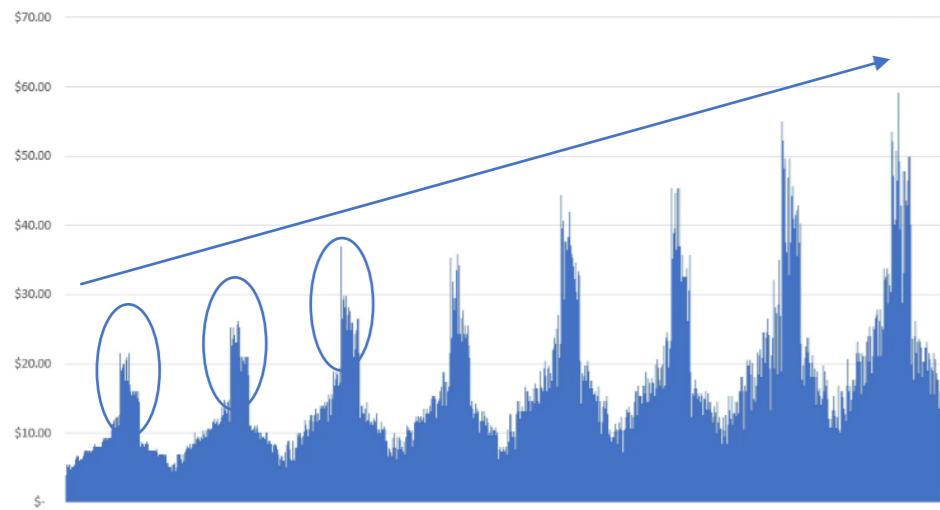
**2.Czynnik B wpływa na czynnik A.** Tu: ludzie zamożniejsi mają lepszy dostęp do wykształcenia i dlatego istnieje związek między zarobkami a wykształceniem.

**3.Jednocześnie A wpływa na B i B na A** Tu: z jednej strony ludzie zamożniejsi mają lepszy dostęp do wykształcenia, ale z drugiej ludzie lepiej wykształceni mają lepsze zarobki.

**4.Istnieje czynnik C niezidentyfikowany w badaniu, który koreluje z A i z B.** Tu: miejsce zamieszkania (lub ambicje) mogą być czynnikiem, który z jednej strony powoduje, że ktoś więcej zarabia, a z drugiej, że ma wyższe wykształcenie

- Identyfikacja, uwagi (cd):
  - Wizualizacje:

Revenue by date



Wartość odstająca – poza  $1,5 * \text{IQR}$

Anomalia – mogą się zawierać w wartościach odstających, ale nie muszą – odstępstwo od typowego zachowania zmiennych.

Czym jest narracja? Mówiąc najprościej, narracja to historia (jak w filmie, książce).

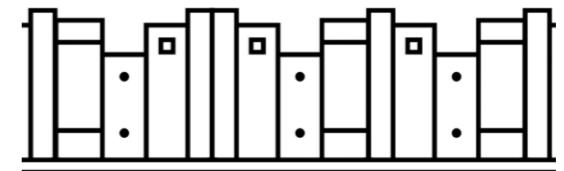
Czym jest narracja oparta na danych ? Narracja oparta na danych to taka, w której wykorzystujemy dane jako część naszej historii, aby rozszerzyć i podkreślić w niej kluczowe punkty.

Pamiętaj, że dane to tylko dane, ale gdy nadamy im znaczenie poprzez umieszczenie ich w kontekście, zamieniają się w informacje (które, mogą służyć uzasadnieniom konkretnych decyzji biznesowych).

Jakie są zatem kluczowe elementy opowieści ?



Ben Howard



Jakie są zatem kluczowe elementy opowieści:

- postacie, osoby lub podmioty, które są zaangażowane w historię (niekoniecznie ludzie),
- sceneria: czas i miejsce gdzie rozgrywa się sceneria,
- jakiś rodzaj konfliktu lub napięcia (to sprawia, że historia jest interesująca, np. problem lub wyzwanie, przed którym stają bohaterowie),
- rozwiązanie lub zakończenie (rozstrzygnięcie to sposób, w jaki konflikt zostaje rozwiązany lub historia się kończy),
- punkt widzenia, czyli perspektywa, z której opowiadana jest historia
- ton i styl, czyli nastrój i sposób, w jaki historia jest przedstawiana (formalnie, nieformalnie, komicznie, dramatycznie).

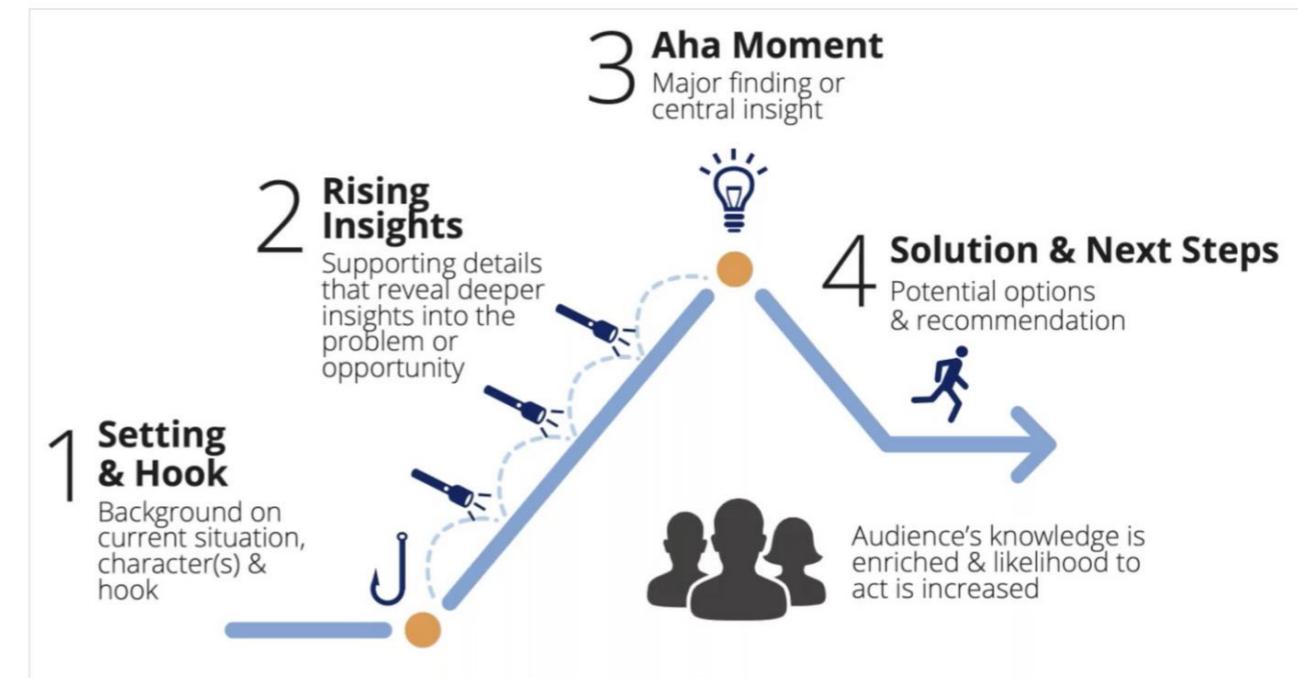


Figure 2. The narrative structure of a data story (Source: [Effective Data Storytelling By Brent Dykes](#))



- „Jestem kierownikiem produkcji w fabryce, w której produkujemy i pakujemy tysiące delikatnych przedmiotów dziennie. Paletyzujemy pudełka, ale niestety niektóre z nich spadają z palet i nasze delikatne produkty ulegają uszkodzeniu. Chcemy wdrożyć system rozpoznawania obrazu, który skanuje każdą paletę przed użyciem i odrzuca te, które mają wady.”

20,000 per day,   stacked onto 370 pallets,  of which 50 break,\* causing approx. 200 broken products per day



Ben Howard

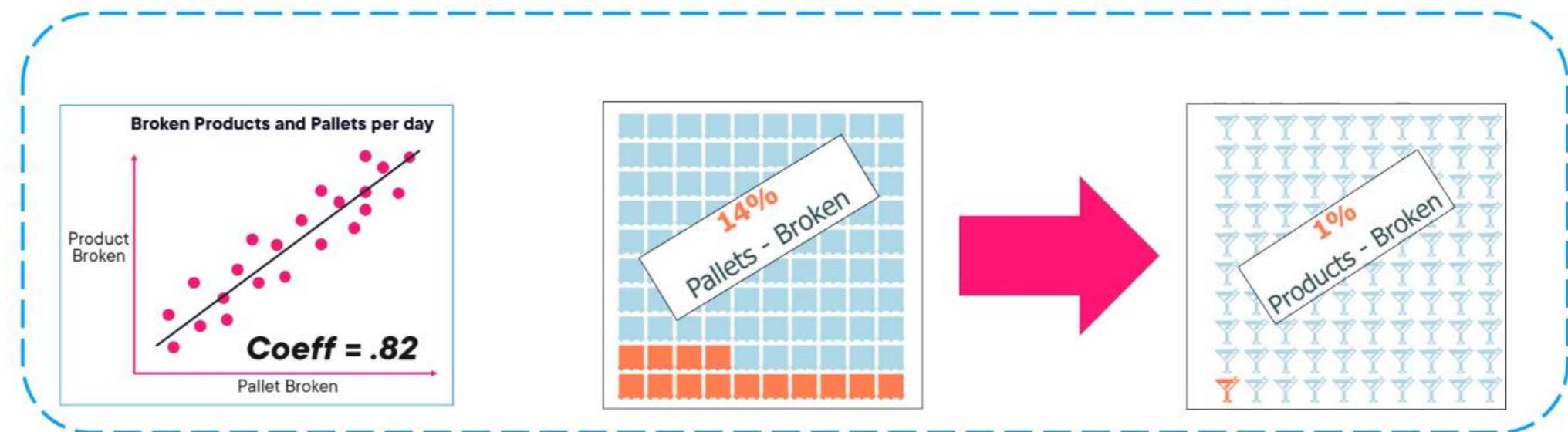
Przekształćmy więc to w narrację opartą na danych (Kierownik przedstawia na formalnym comiesięcznym spotkaniu):

„Produkujemy 20 000 produktów dziennie, które są układane na 370 paletach, a z 370 palet mamy problemy z 50 z nich, co powoduje, że jedno lub dwa pudełka spadają na podłogę pakowania, powodując uszkodzenia i przeróbki. Łącznie tracimy około 200 produktów dziennie.

Przez okres miesiąca mierzyłem liczbę uszkodzonych produktów i liczbę uszkodzonych palet, a następnie obliczyłem współczynnik korelacji, który wyniósł 0,82, co wskazuje na silną, dodatnią korelację. Następnie wykreśliłem wartości na wykresie punktowym, a wyniki można zobaczyć na ekranie. Przeanalizowałem również średnią liczbę palet z vadami, która wyniosła średnio 14% całkowitego zapasu palet. Następnie przeprowadziłem analizę przyczyn źródłowych, a te 14% palet bezpośrednio spowodowało uszkodzenie 1% naszych produktów. Z tego możemy wywnioskować, że rozwiązanie problemu uszkodzonych palet rozwiąże problem uszkodzonych produktów.”

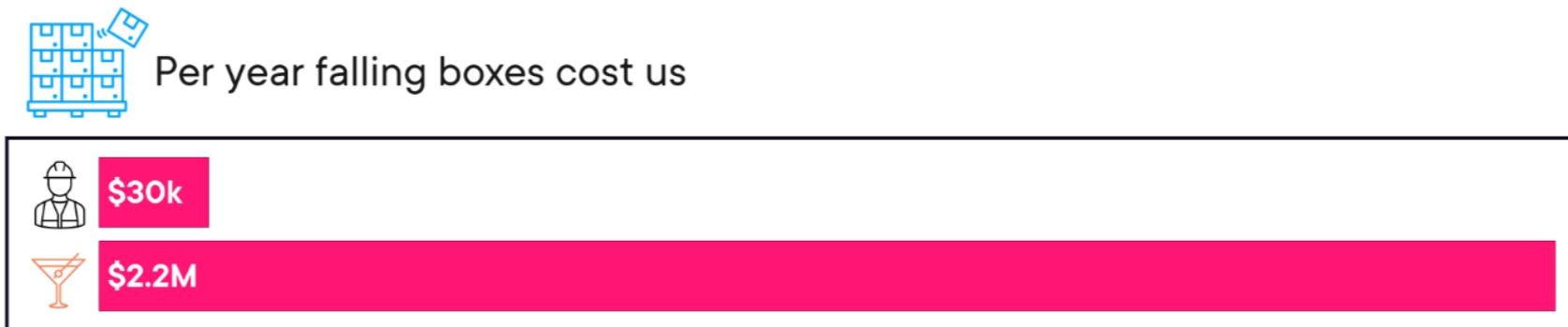
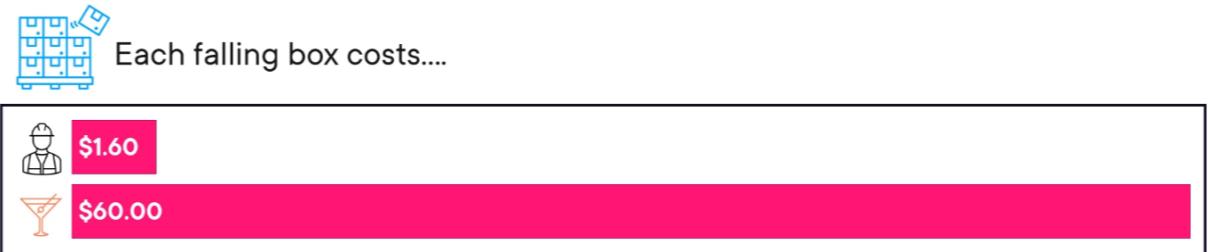


Ben Howard



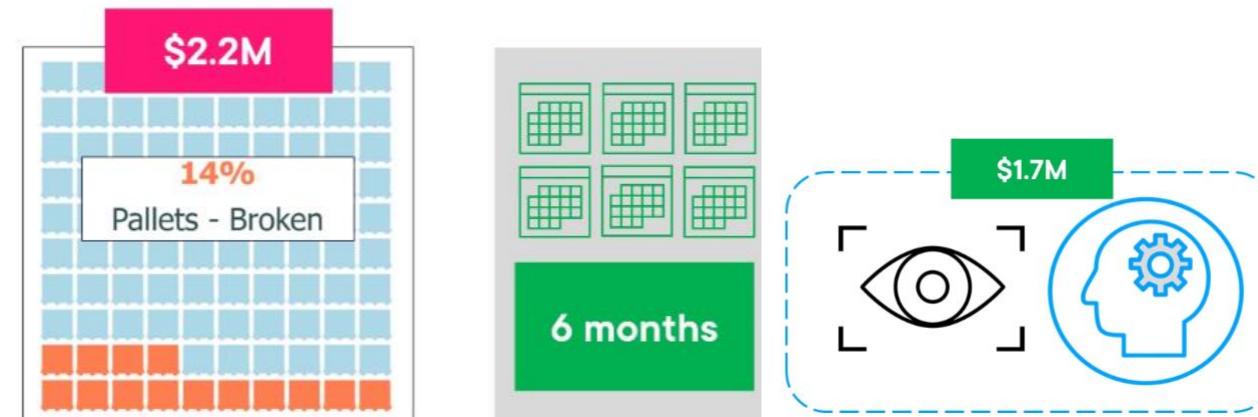
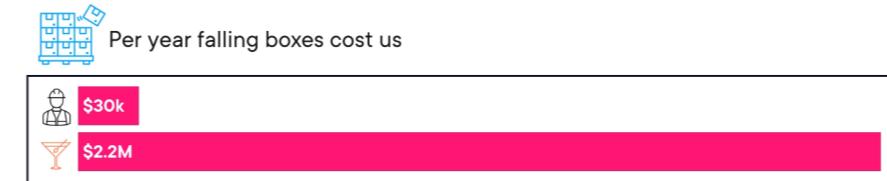
## Teraz problem jest zrozumiany, można dokonać analizy czy opłaca się to naprawiać ?

- „Każda spadająca skrzynka kosztuje, a koszty dzielą się na dwa obszary.
- Pierwszym z nich jest robocizna. Naprawa palety zajmuje około 5 minut i kosztuje 1,60 USD, co samo w sobie nie jest dużą kwotą.
- Prawdziwym kosztem jest produkt. Każdy produkt sprzedawany jest za 30,00 USD, a średnio dwa produkty są uszkodzone w każdym pudełku, co daje 60,00 USD za pudełko.
- Patrząc na to w ujęciu rocznym, koszt robocizny wynosi 30 000 USD, ale wartość produktu to **2,2 miliona USD**, co myślę, że wszyscy zgodzimy się, że jest znaczące.



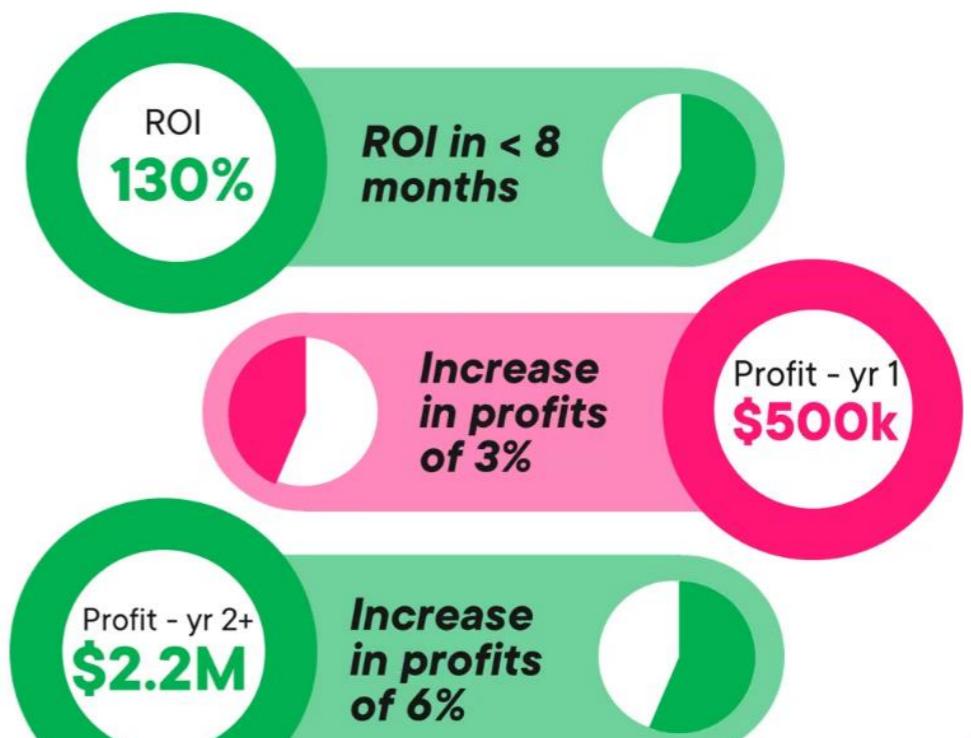
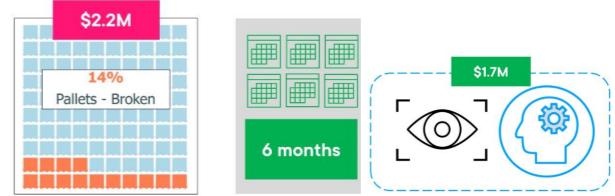
Ben Howard

- Chciałbym teraz przedstawić proponowane przez nas rozwiązanie:
- Po przeprowadzeniu analizy przyczyn źródłowych wiemy, że 14% naszych palet ma wady, co jest bezpośrednią przyczyną utraty wartości **2,2 miliona dolarów rocznie**.
- Koszt wdrożenia i przeszkołenia naszego rozwiązania do rozpoznawania obrazów wynosi **1,7 miliona dolarów i można to osiągnąć w ciągu 6 miesięcy**.



Ben Howard

- Przeanalizowałem te finanse z naszym dyrektorem finansowym, a ona przedstawiła następujące dane.
- **ROI**, czyli zwrot z inwestycji, wynosi 130%, a **zwrot z inwestycji nastąpi w mniej niż 8 miesięcy**.
- Powinno to doprowadzić do zysku w pierwszym roku w wysokości 50 000 USD, co odpowiada wzrostowi o 3%, ale ponieważ ROI jest krótszy niż 1 rok, zyski w latach 2+ wynoszą 2,2 miliona USD, czyli wzrost zysków o 6%.
- Jak widać, uważam, że wdrożenie systemu rozpoznawania obrazów i produktu uczenia maszynowego będzie wartościową inwestycją dla naszej firmy. Dziękuję za wysłuchanie. Teraz chętnie odpowiem na pytania. [...]"



Ben Howard

1. Twórz narracje danych z przemyślanym przesłaniem
2. Pozycjonuj elementy narracji danych
3. Rozumiej swoją publiczność, uwzględnij poziomy emocji ([trzy poziomy emocji, wg. Norman](#)):

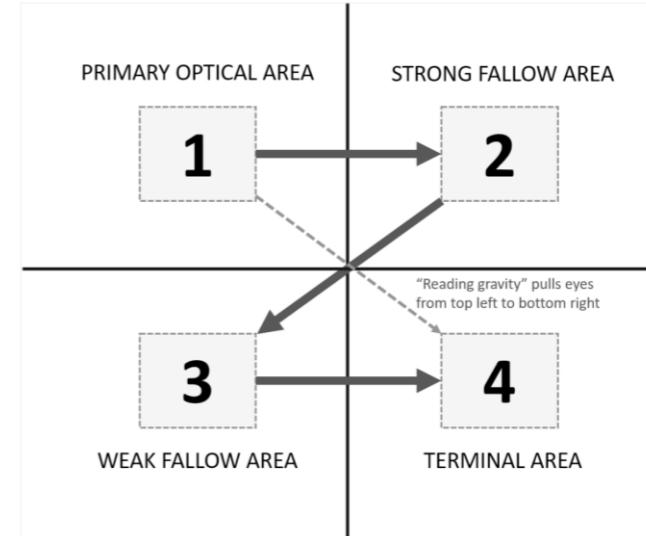
Don Norman argumentuje, że użytkownicy końcowi przetwarzają wszystkie obiekty na trzech poziomach emocji – wisceralnym, behawioralnym i refleksyjnym:

**Poziom emocjonalny (visceral):** W pierwszej sekundzie widzenia wykresu, użytkownicy oceniają wykres na podstawie jego wyglądu. Jest to poziom przetwarzania, który odnosi się do zakorzenionych i automatycznych cech ludzkich emocji. Estetyka wykresu jest kluczowa, aby utrzymać uwagę publiczności na narracji danych.

**Poziom behawioralny (behavioral):** Następnie użytkownik ocenia wizualizację na podstawie jej użyteczności, co określa się mianem poziomu behawioralnego przetwarzania. Innymi słowy, publiczność musi być w stanie zrozumieć punkt autora z wizualizacji.

**Poziom refleksyjny (reflective):** Po interakcji z wizualizacją, użytkownicy zastanawiają się nad swoim doświadczeniem. Na przykład użytkownik prawdopodobnie ponownie skorzysta z pulpitu danych, jeśli uzna swoje doświadczenie za przyjemne po refleksji.

4. Użyj struktury narracyjnej w opowieściach danych
5. Dostosuj przekaz do formy przekazu (czy detaliczny wykres będzie czytelny?)
6. Odgrąć wizualizacje (less is more)
7. Opowiadaj historie w kawałkach ...



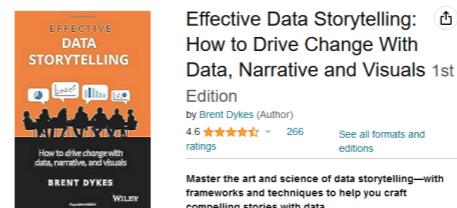
Gutenberg principle: design principle that describes the general movement of the eyes when looking at a design in which elements are evenly distributed. It's also known as the Gutenberg Rule or the Z pattern of processing

## Why we love (or hate) everyday things

FORBES > LEADERSHIP > ENTREPRENEURS

### Why Data Storytellers Will Define The Next Decade Of Data

Brent Dykes Contributor @  
I write about how to drive more value with data and analytics.



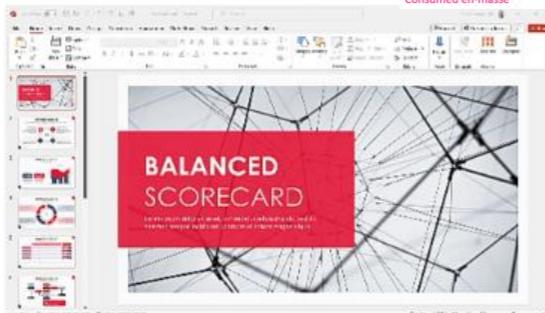
## Data Insights -> Data Driven Narrative -> Data Storytelling

- Komunikowanie informacji z danych - formy:

Multiple pages of information  
Typically written in A4 size  
Both electronic and physical  
Fully explores a single topic  
Includes summaries, an appendix, chapters  
Textual with supporting images  
Static data  
Reader focus hours to days  
Consumed individually



Multiple pages of information  
Both electronic and physical  
Explores a single area  
Both text and images  
Displays static data  
Audience focus; 10 mins+  
Consumed en-masse



Multiple pages of information  
Both electronic and physical  
Exploring a single area  
Both text and images  
Displays static data  
Audience focus; 10 mins+  
Consumed en-masse



Single page of information  
A5 to large display boards  
Often physical  
Explain complex information in easily digestible pieces  
Visual imagery with textual explanation  
Display static data  
Reader focus < 5 mins  
Centered on a single subject  
Consumed individually

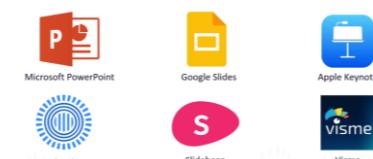


Single page of information  
Electronic only  
Covers multiple areas  
Based around multiple visuals  
Displays live data  
Viewer focus; seconds to minutes  
Interactive, ingress point for further data analysis

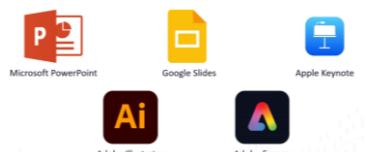
Report



Presentation



Infographic



Dashboard



## Wizualizacja danych

Dashboard (pulpit nawigacyjny/kokpit managerski) i Scorecard (karta wyników) - oba te pojęcia agregują dane z różnych jednostek biznesowych i dają czytelnikowi możliwość monitorowania wydajności danych wskaźników.

Zarówno dashboardy, jak i karty wyników stanowią kulminację analityki biznesowej.

Interfejs pulpitu nawigacyjnego lub karty wyników ułatwia użytkownikom szybkie znalezienie, analizowanie i eksplorowanie informacji, których potrzebują do wykonywania swojej pracy.

Zarówno pulpity nawigacyjne, jak i karty wyników są narzędziami analitycznymi, które pozwalają skupić się na pomiarach ważnych dla firmy.

Niektórzy ludzie używają terminów dashboard i scorecard zamiennie, podczas gdy inni używają tych terminów w odniesieniu do różnych typów aplikacji analitycznych. Zarówno dashboardy i karty wyników to po prostu różne rodzaje wizualnych mechanizmów wyświetlania w ramach systemu zarządzania wydajnością (danego procesu), które przekazują krytyczne (na pierwszy rzut oka) informacje o wydajności.

Podstawowa różnica polega na tym, że dashboardy mają tendencję do monitorowania wydajności procesów operacyjnych (zarządzanie niższego szczebla) podczas gdy karty wyników mają tendencję do wykreślania postępów w realizacji celów taktycznych i strategicznych (top management).

Inni interesariusze – pamiętaj jednak – dostosuj do sytuacji w której znajdujesz się w organizacji, niezależnie od nomenklatury.

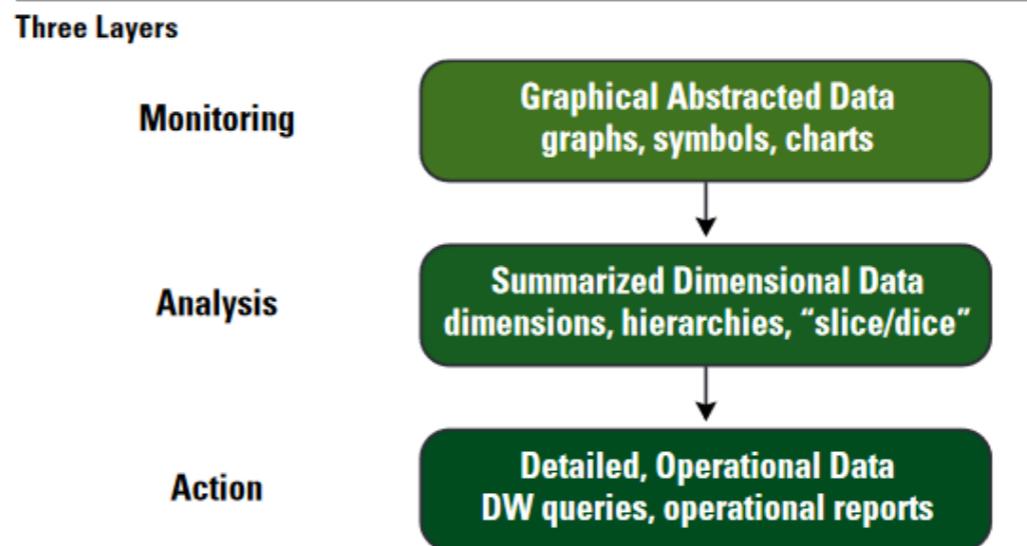


Figure 2. A performance management system provides access to three layers of data.

## Wizualizacja danych

### Dashboard

**Cel:** Dashboardy (pulpity nawigacyjne) są zaprojektowane do monitorowania i przedstawiania wydajności operacyjnej na bieżąco. Ich głównym zadaniem jest śledzenie procesów i operacji w czasie rzeczywistym lub na krótkich interwałach czasowych.

#### Charakterystyka:

**Wizualizacje:** Dashboardy często wykorzystują dynamiczne wizualizacje, takie jak wykresy, grafy, mapy ciepła i tabele.

**Interaktywność:** Zazwyczaj oferują wysoki poziom interaktywności, pozwalając użytkownikom na filtrowanie, sortowanie i zagłębianie się w dane.

**Monitorowanie:** Skupią się na kluczowych wskaźnikach efektywności operacyjnej (KPIs), które są ważne dla bieżących operacji.

#### Dashboards versus Scorecards

	Dashboard	Scorecard
Purpose	Measures performance	Charts progress
Users	Managers, staff	Executives, managers, staff
Updates	Real-time to right-time	Periodic snapshots
Data	Events	Summaries
Top-level Display	Charts and tables	Symbols and icons

Figure 3. Dashboards and scorecards are visual display mechanisms—the monitoring layer in a performance dashboard.  
A good performance dashboard should support both types of displays.

### Scorecard

**Cel:** Karty wyników (scorecards) są używane do śledzenia postępów w realizacji celów strategicznych i taktycznych. Służą one przede wszystkim do oceny długoterminowych trendów i wyników w stosunku do założonych celów.

#### Charakterystyka:

**Metodyka:** Często oparte na podejściu *Balanced Scorecard*, która pomaga w mierzeniu sukcesów na różnych poziomach organizacji (Finance, Customer, Internal, Learning Growth).

**Symbolika:** Używają graficznych symboli, takich jak strzałki, kolory i ikony, aby pokazać kierunek trendów (np. poprawa, pogorszenie) i osiągnięcie celów.

**Kontekst strategiczny:** Koncentrują się na długoterminowych celach i strategiach firmy, mniej na codziennych operacjach. Zaprojektowane do szybkiego przekazu informacji o tym, czy firma jest na dobrej drodze do realizacji swoich strategicznych celów.

## A wg niektórych badań...

### Types of Users

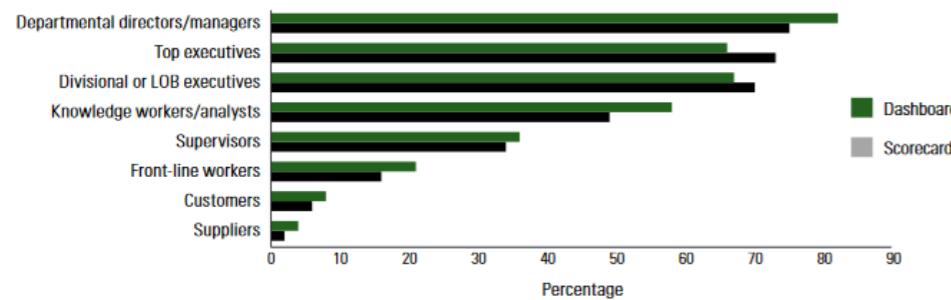


Figure 10. Based on 299 and 199 respondents, respectively, who have deployed or are about to deploy either dashboards or scorecards.

Eckerson, W. (2006). Deploying dashboards and scorecards. *The Data Warehouse Institute*, 1-24.  
 Zingde, S., & Shroff, N. (2020). The Role of Dashboards in Business Decision Making and Performance Management. *A Road Map to Future Business; Institute of Management, Nirma University: Ahmedabad, India*, 227.

### Does Your Group Support Both a Dashboard and Scorecard?

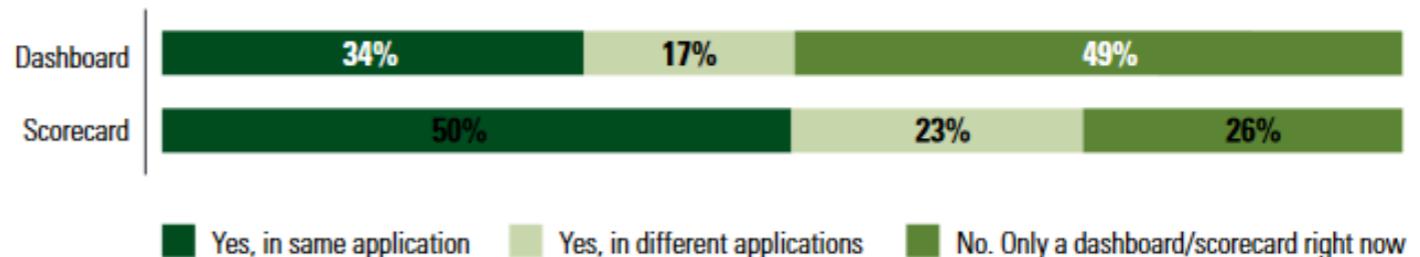
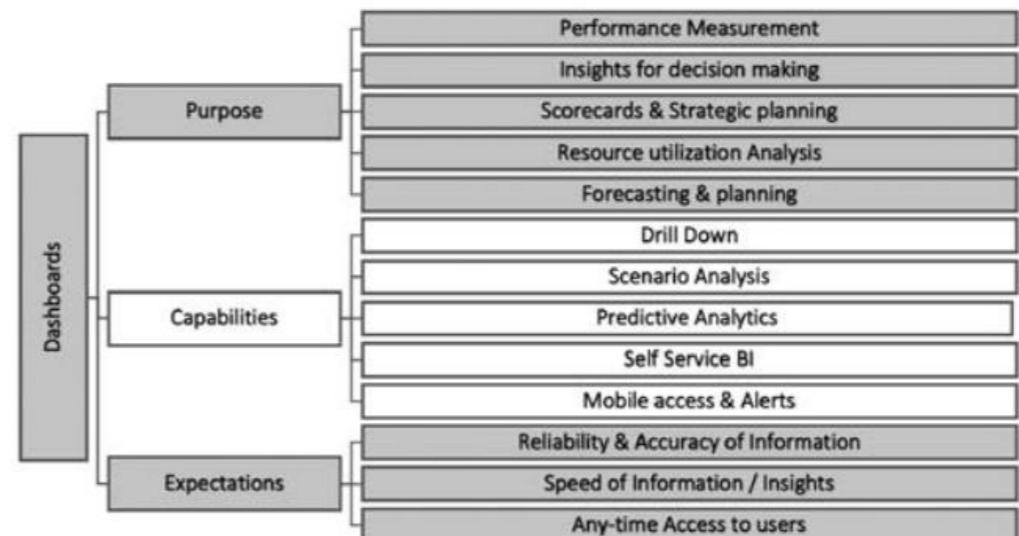


Figure 7. Based on 299 and 199 respondents who have deployed or are about to deploy dashboards or scorecards, respectively.

Figure 1: Dashboards - Purpose, Capabilities and Expectations

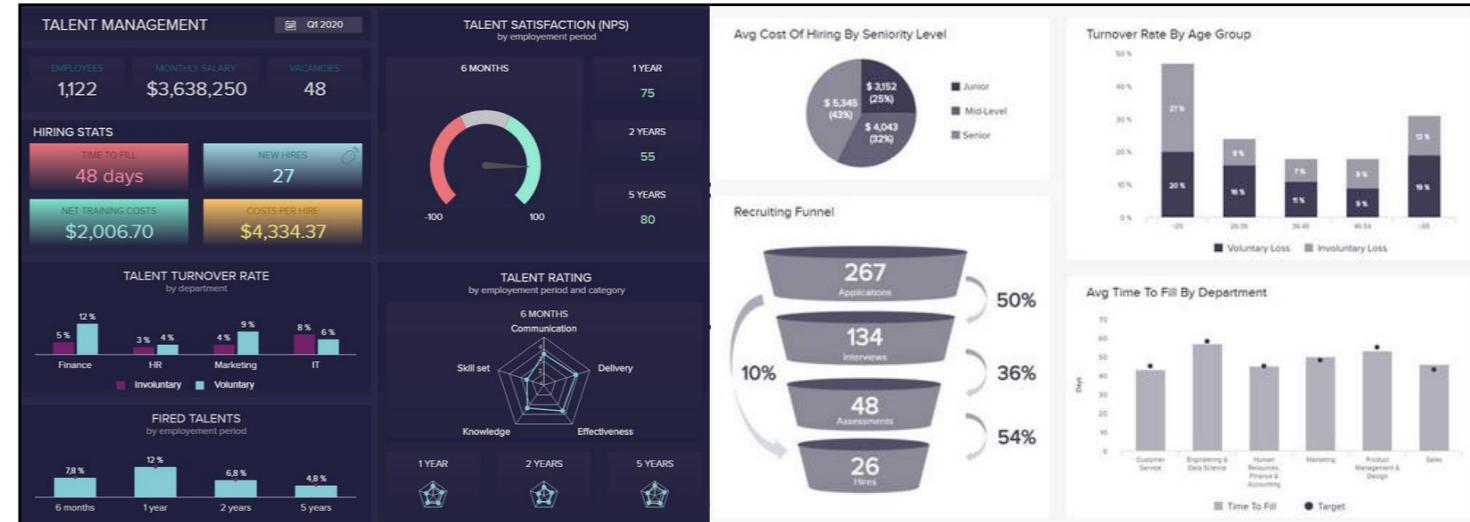


Noonpakdee and Khunkornsiri, (2018) proposed framework comprising 4 main components - business operation, attribute list, visualization, and dashboard capabilities. Researcher consolidated various learnings to provide a quick overview in above Figure 1: Dashboards - Purpose, Capabilities and Expectations.

## Scorecards

## Dashboards

# Use both in synergy!



Scorecard	
Performance Management	
KPI (Metric & Target)	
Progress (Current Value vs. The Target)	
Periodically (Daily/Weekly/Monthly/Quarterly)	
Long Term Goal	
Companies' Policies	
Strategic	
Top Management	
Trends & Changes In Business Activity Over Period Of Time	
Summarized/Consolidated	

Dashboard	
Performance Monitoring	
Performance Metric	
Performance	
Real Time Basis	
Short Term Goal	
Daily Operations	
Tactical	
Individual Managers	
Snapshot Of Business Performance	
Real Time Data Obtained	

<https://www.datapine.com>

**Automatyzacja i skalowalność:** Raporty BI powinny być łatwe w utrzymaniu i aktualizacji. Automatyzacja procesu zgromadzenia danych, ich przetwarzania oraz dystrybucji raportu jest kluczowa, zwłaszcza w dynamicznie zmieniających się środowiskach. Skalowalność rozwiązania pozwala na rozbudowę i modyfikację raportów w miarę ewolucji potrzeb organizacji.

- **Automatyzacja:** Maksymalizacja automatyzacji procesów przetwarzania i analizy danych, aby zminimalizować potrzebę interwencji człowieka i zwiększyć efektywność operacyjną.

*Automated Data Pipelines: Tworzenie zautomatyzowanych łańcuchów przetwarzania danych, które integrują dane z różnych źródeł, przetwarzają je i dostarczają wyniki w formie gotowych analiz i raportów.*

#### Zasady:

- **Modularność:** Projektowanie systemów w sposób modularny, co pozwala na łatwe dodawanie, usuwanie lub modyfikowanie komponentów systemu **bez zakłócania całości**.
- **Skalowalność i efektywność:** Stałe monitorowanie i ocena wydajności systemów BI w celu identyfikacji obszarów do optymalizacji i zapewnienia ciągłej poprawy wydajności.
  - **Load Balancing (Balansowanie obciążenia):** Dystrybucja pracy między różne komponenty systemu w celu optymalnego wykorzystania zasobów i zapobiegania przeciążeniom.
  - **Elastyczność:** Tworzenie systemów zdolnych do dostosowywania się do zmieniających się obciążień i wymagań, zarówno pod względem przetwarzania danych, jak i ich przechowywania.
  - **Orkiestracja:** Koordynacja i zarządzanie automatycznymi procesami i zadaniami w ramach systemów BI, cykliczne wykonywanie codziennych zadań związanych z łączeniem i przetwarzaniem danych czy nauką algorytmów uczenia maszynowego. Umożliwi też łatwą obserwację, edycję, rozwój i harmonogramowanie, a w przypadku wystąpienia błędów i niepowodzeń, zapewniającą, że wszystkie składowe działają harmonijnie (np. Azure Data Factory, ETL, Data Pipeline Implementation, Real-time data pipelines, Batch Processing Systems).

**Automatyzacja i skalowalność:** Raporty BI powinny być łatwe w utrzymaniu i aktualizacji. Automatyzacja procesu zgromadzenia danych, ich przetwarzania oraz dystrybucji raportu jest kluczowa, zwłaszcza w dynamicznie zmieniających się środowiskach. Skalowalność rozwiązania pozwala na rozbudowę i modyfikację raportów w miarę ewolucji potrzeb organizacji

## Kierunek -> **Hiper automatyzacja** (Według Gartnera RPA (ang. Robotic Process Automation, RPA) wzbogacone o AI i ML staje się rdzeniem technologii hiperautomatyzacji):

*Automatyzacja koncentruje się na wąskim zestawie zadań, często opierając się na technologii RPA albo łącząc dwie technologie w jedno narzędzie, które automatyzuje i optymalizuje istniejące zadania bądź procesy.*

*Hiperautomatyzacja nie jest narzędziem - raczej ujednoliczoną strategią lub inicjatywą przedsiębiorstwa, której ostatecznym celem jest tworzenie i optymalizacja kompleksowych procesów, tworzących nowe możliwości biznesowe. To koncepcja, która pozwala na cyfrową transformację całego przedsiębiorstwa.*

*Kroki:*

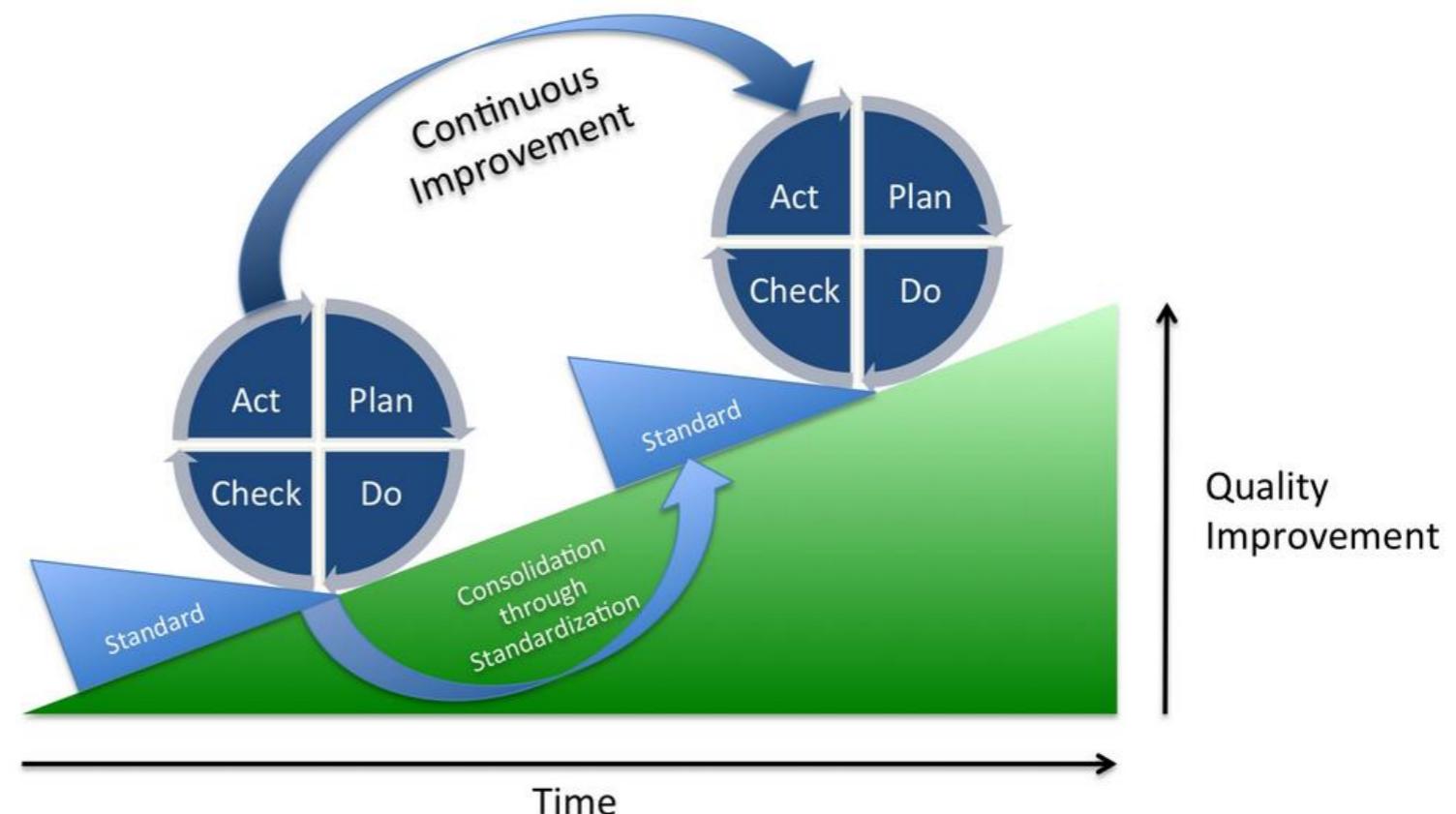
- **Optymalizacja procesów** (przyjrzeć się procesom w firmie. Automatyzowanie procesów, które są zbędne, niepotrzebnie skomplikowane lub przestarzałe jest nieprawidłowym działaniem. **Najpierw powinniśmy zoptymalizować wszystkie ścieżki procesowe, a dopiero później wdrażać automatyzację**)
- **Automatyzacja procesów** (Aby wdrożyć strategię hiperautomatyzacji, **przedsiębiorstwo potrzebuje solidnych podstaw automatyzacji**. Istotny jest **dobór zestawu konkretnych technologii**, który może obejmować i zwykle obejmuje narzędzie RPA do automatyzowania podstawowych zadań, rozwiązania automatyzacyjne IT dla hurtowni danych oraz kilka innych, **które są niezbędne do obsługi różnych zespołów i działów w danej organizacji**.)
- **Orkiestracja, czyli dostrojenie i dopasowanie wybranych narzędzi** (Ważna jest optymalizacja działania robotów i wprowadzenie sztucznej inteligencji (AI) oraz uczenia maszynowego (ML) pod kątem wymagań biznesowych. Zastosowanie zaawansowanych technologii, takich jak przetwarzanie języka naturalnego (NLP – ang. natural language processing), optyczne rozpoznawanie znaków (OCR – ang. optical character recognition), zaawansowana analityka i cyfrowy bliźniak organizacji (DTO - Digital Twin Organization) w celu tworzenia innowacyjnych nowych procesów.)

**Regularne przeglądy i aktualizacje:** Świat biznesu szybko się zmienia, więc raporty BI również powinny być regularnie przeglądane i aktualizowane, aby zapewnić, że nadal odpowiadają na aktualne potrzeby i wyzwania biznesowe.

## Zasady:

- 1. Zaplanowane przeglądy:** Ustanowienie harmonogramu regularnych przeglądów raportów, np. co kwartał, co pół roku lub rocznie, w zależności od charakteru biznesu i szybkości zmian w branży.
- 2. Uwzględnienie feedbacku użytkowników:** Systematyczne zbieranie i analiza opinii użytkowników raportów BI, aby dowiedzieć się, które aspekty raportów są najbardziej i najmniej przydatne, oraz jakie nowe potrzeby się pojawiły.
- 3. Monitorowanie wskaźników KPI:** Stale monitorowanie kluczowych wskaźników wydajności (KPIs), które raporty mają śledzić, oraz dostosowywanie ich do aktualnych celów biznesowych.

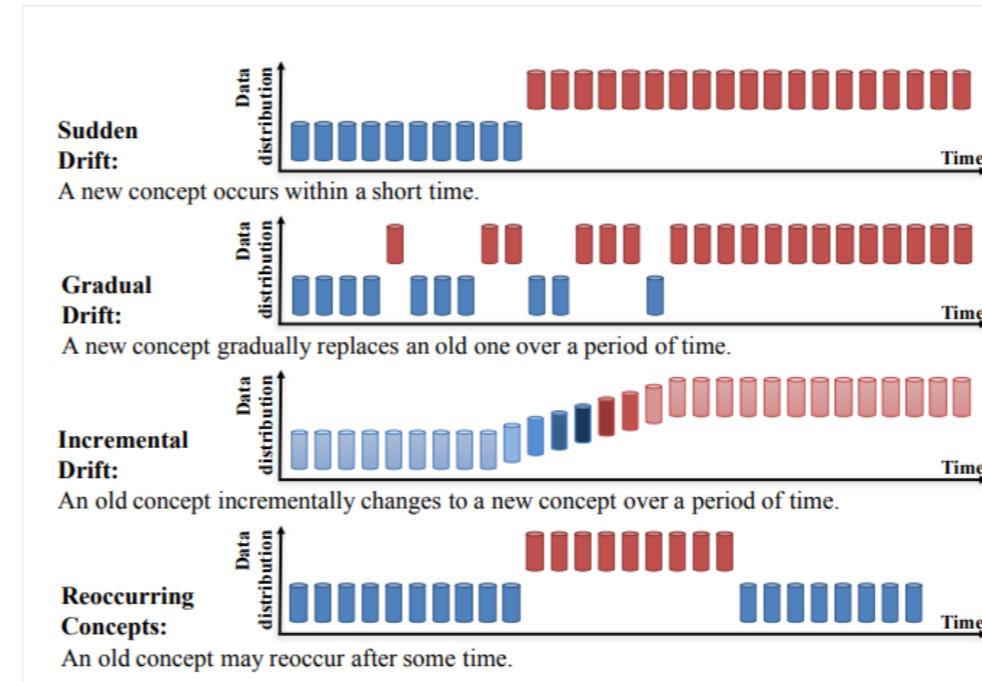
**Cycle of Continuous Improvement (Cykl ciągłego doskonalenia):**  
Proces ciągłej optymalizacji organizacji (w tym także raportów BI), w którym regularne przeglądy są kluczowym elementem.



**Data Drift (Dryf danych):** Zjawisko zmiany charakterystyki danych na przestrzeni czasu, co może wpływać na skuteczność i dokładność raportów BI.

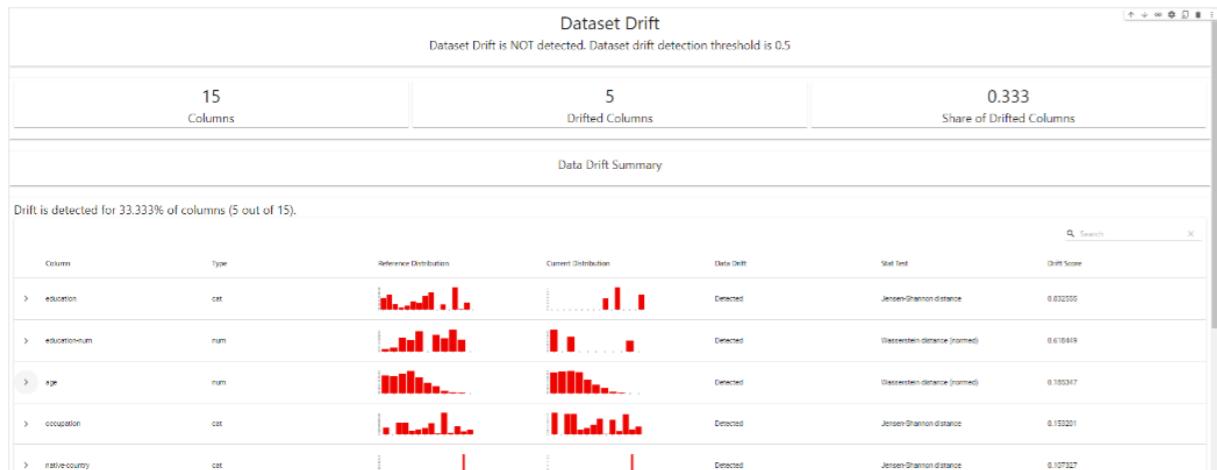
"Dryf" to termin używany w uczeniu maszynowym do opisania, w jaki sposób wydajność modelu uczenia maszynowego w produkcji powoli pogarsza się z czasem. Może się to zdarzyć z wielu powodów, takich jak zmiany w rozkładzie danych wejściowych w czasie lub zmiana relacji między danymi wejściowymi (x) a pożądanym celem (y).

Dryf koncepcji, znany również jako dryf modelu, występuje, gdy zadanie, do którego model został zaprojektowany, zmienia się w czasie. Wyobraźmy sobie na przykład, że model uczenia maszynowego został przeszkolony do wykrywania spamu na podstawie treści wiadomości e-mail. Jeśli rodzaje spamu, które ludzie otrzymują, znacznie się zmieniają, model może nie być już w stanie dokładnie wykrywać spamu.

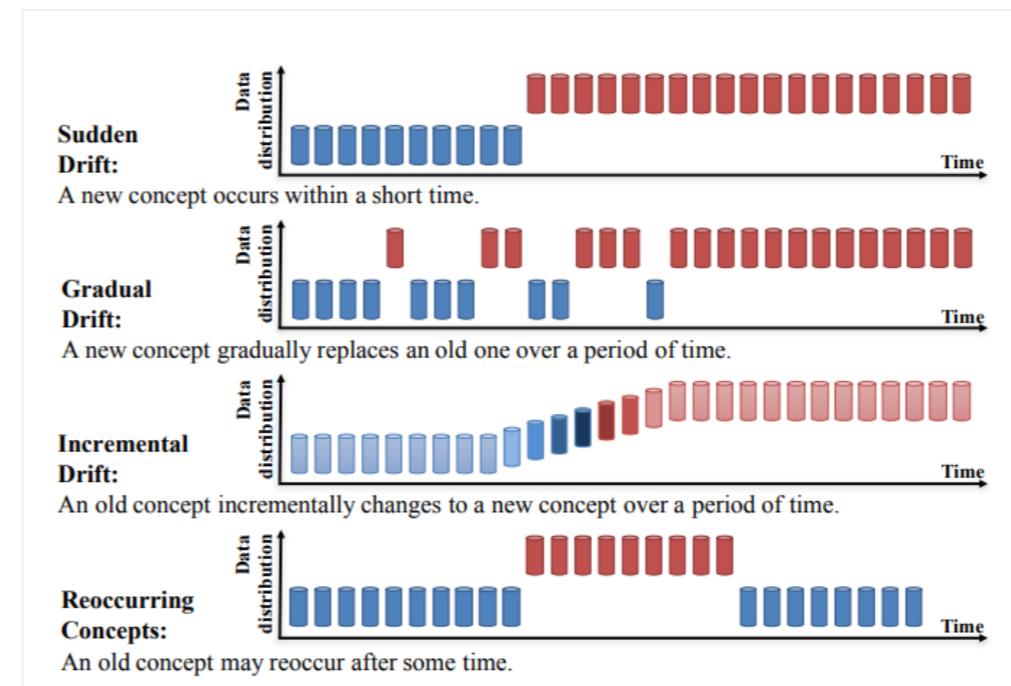


**Data Drift (Dryf danych):** Zjawisko zmiany charakterystyki danych na przestrzeni czasu, co może wpływać na skuteczność i dokładność raportów BI.

"Dryf" to termin używany w uczeniu maszynowym do opisania, w jaki sposób wydajność modelu uczenia maszynowego w produkcji powoli pogarsza się z czasem. Może się to zdarzyć z wielu powodów, takich jak zmiany w rozkładzie danych wejściowych w czasie lub zmiana relacji między danymi wejściowymi (x) a pożądanym celem (y).



Drift Detection Dashboard - created using EvidentlyAI

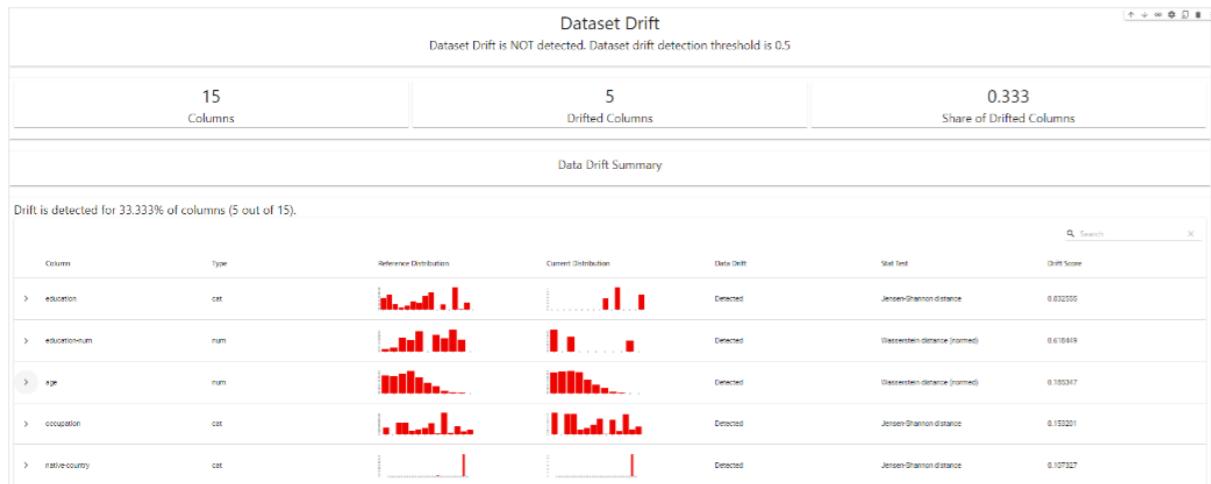


Learning under Concept Drift: A Review  
 Jie Lu, Fellow, IEEE, Anjin Liu, Member, IEEE, Fan Dong, Feng Gu, Jo~ao Gama, and Guangquan Zhang

## Regularne przeglądy i aktualizacje

**Data Drift (Dryf danych):** Zjawisko zmiany charakterystyki danych na przestrzeni czasu, co może wpływać na skuteczność i dokładność raportów BI.

"Dryf" to termin używany w uczeniu maszynowym do opisania, w jaki sposób wydajność modelu uczenia maszynowego w produkcji powoli pogarsza się z czasem. Może się to zdarzyć z wielu powodów, takich jak zmiany w rozkładzie danych wejściowych w czasie lub zmiana relacji między danymi wejściowymi (x) a pożądanym celem (y).



Drift Detection Dashboard - created using EvidentlyAI

## Kolmogorov-Smirnov (K-S) test Population Stability Index Page-Hinkley method

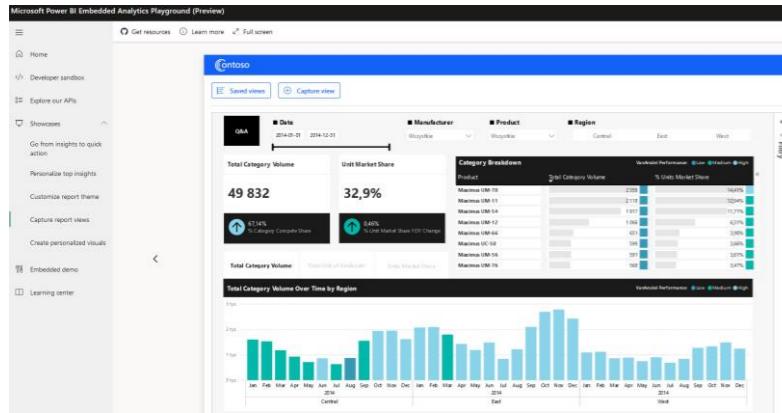
**more:**

**<https://www.datacamp.com/tutorial/understanding-data-drift-model-drift>**

Learning under Concept Drift: A Review

Jie Lu, Fellow, IEEE, Anjin Liu, Member, IEEE, Fan Dong, Feng Gu, Jo˜ao Gama, and Guangquan Zhang

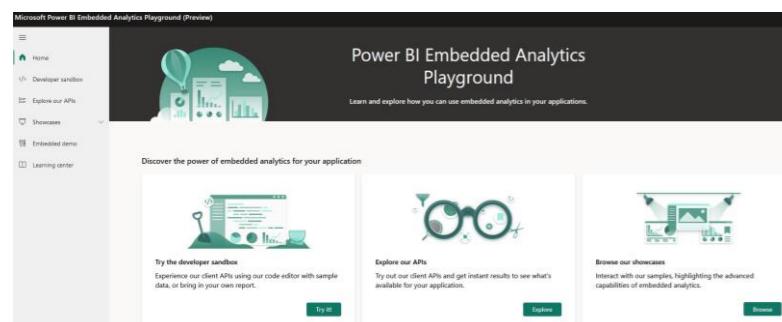
## Regularne przeglądy i aktualizacje



**Sandboxing:** Umożliwienie użytkownikom tworzenia własnych raportów i analiz w kontrolowanym środowisku (tzw. piaskownica), co może prowadzić do identyfikacji potrzeb aktualizacji oficjalnych raportów.

**Eksperymentowanie z danymi:** pozwala użytkownikom na swobodne eksperymentowanie z danymi bez obawy o wpływ na produkcyjne środowisko analityczne. Mogą oni importować własne dane, testować różne wizualizacje i modele oraz tworzyć nowe wskaźniki i miary.

**Prototypowanie i testowanie:** może być używany do prototypowania i testowania nowych dashboardów, raportów i analiz przed ich wdrożeniem w środowisku produkcyjnym. Pozwala to na uzyskanie informacji zwrotnej od użytkowników i udoskonalenie analityki przed jej udostępnieniem szerszej publiczności.



**Uczenie się i rozwój:** może służyć do nauki nowych technik analitycznych i rozwijania umiejętności analitycznych. Użytkownicy mogą brać udział w samouczkach, ćwiczeniach i innych zasobach edukacyjnych dostępnych w sandboxie.

# Podsumowanie – złote zasady

Visuals are processed  
**60.000 times faster than  
text!**



Target audience



Keep it simple



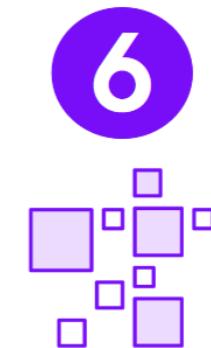
Consistency



Remove “noise”



Color selection



Highlight key data

# Ćwiczenie 10 – wizualizacja (1)

**Cel:** Stworzenie interaktywnego dashboardu w Power BI, który umożliwi użytkownikom szybki wgląd w kluczowe wskaźniki efektywności (KPIs) dla wybranej firmy lub działu.

**Dane:** Zasymuluj zestaw danych, który zawiera informacje sprzedażowe firmy z ostatniego roku, w tym dane na temat sprzedaży, kosztów, zysków i klientów (możesz użyć danych z poprzednich zadań)..

## Kroki do wykonania:

**1. Import danych:** Zaimportuj dane do Power BI.

**2. Modelowanie danych:** Utwórz odpowiednie relacje w modelu danych.

**3. Tworzenie KPIs:** Zdefiniuj i oblicz KPIs takie jak całkowita sprzedaż, marża zysku, nowi klienci, wskaźnik zadowolenia klientów.

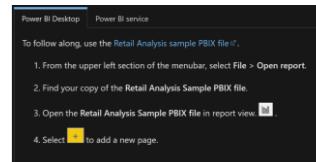
### 4. Projektowanie dashboardu:

1. Użyj różnych form wizualizacji (wykresy, mapy ciepła, histogramy).
2. Zaimplementuj filtry i slicery, aby użytkownicy mogli filtrować dane na podstawie różnych kryteriów (np. region, produkt).
3. Dodaj funkcje interaktywne, takie jak drill-down, aby umożliwić użytkownikom szczegółową analizę danych.

**5. Testowanie i optymalizacja:** Sprawdź funkcjonalność i estetykę dashboardu, upewniając się, że jest intuicyjny i efektywny.

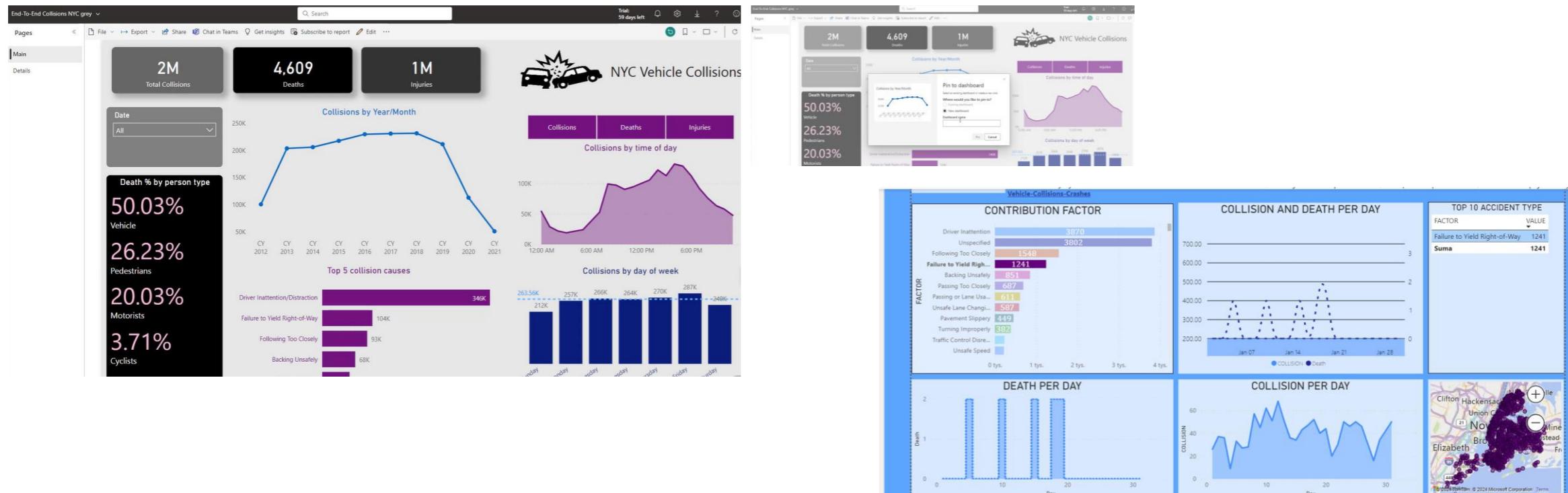
**Wymagania końcowe:** Dashboard powinien być funkcjonalny, estetycznie wykonany i łatwy w obsłudze. Proszę objąć logikę biznesową stojącą za wybranymi KPIs oraz zademonstrować funkcjonalność dashboardu. Proszę opisać wykonane kroki i przesłać je (wraz z rezultatami ćwiczenia 11) Prowadzącemu do dn. **22.05.2024**

Przykład: <https://learn.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-kpi?tabs=powerbi-desktop>



# Ćwiczenie 11 – wizualizacja (2)

Pobierz dane ze strony: [https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about\\_data](https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data) na temat kolizji samochodowych w NYC. Przygotuj interaktywny raport z możliwością podglądu szczegółów na wzór:



Opisz wykonane kroki (sprawozdanie) i prześlij raporty (wraz z sprawozdaniem) Prowadzącemu wraz z ćwiczeniem 10 do dn. **22.05.2024**

# „on-premise” i "cloud" BI Big Picture

Graphics: Ben Sullins

# Kontext

"on-premise" i "cloud" to dwa różne modele dostarczania i hostowania aplikacji oraz infrastruktury

## On-Premise (Lokalnie)

### 1. Lokalna infrastruktura:

W przypadku rozwiązań on-premise infrastruktura, w tym serwery, sieci, pamięć masowa, znajduje się fizycznie w siedzibie firmy lub w centrum danych, której jest właścicielem lub zarządzającym.

### 2. Pełna kontrola:

Organizacja ma pełną kontrolę nad infrastrukturą oraz oprogramowaniem, co pozwala na dostosowanie do własnych potrzeb i regulacji.

### 3. Koszty inwestycyjne:

Wymaga inwestycji w zakup i utrzymanie sprzętu oraz oprogramowania, co może być kosztowne zarówno na początku, jak i w dłuższej perspektywie.

### 4. Wymagana kadra:

Wymaga utrzymania zasobów IT w firmie lub zatrudnienia zewnętrznej firmy do zarządzania infrastrukturą.

## Cloud (Chmura)

### 1. Hostowane przez dostawcę:

Infrastruktura jest hostowana przez dostawcę usług chmurowych, taki jak Amazon Web Services (AWS), Microsoft Azure lub Google Cloud Platform (GCP).

### 2. Elastyczność i skalowalność:

Usługi w chmurze oferują elastyczność i skalowalność, co oznacza, że można łatwo dostosować zasoby do zmieniających się potrzeb.

### 3. Modele cenowe:

Oferowane są różne modele cenowe, takie jak płatność za zużyte zasoby (pay-as-you-go) lub abonamenty okresowe, co może być bardziej elastyczne niż tradycyjne modele licencyjne.

### 4. Uproszczone zarządzanie:

Zarządzanie infrastrukturą jest często ułatwione dzięki narzędziom i usługom dostępnym w chmurze, co może zmniejszyć obciążenie dla zespołu IT.

### 5. Dostępność i wydajność:

Dostawcy chmur zazwyczaj oferują wysoką dostępność i wydajność poprzez dystrybucję danych na wiele regionów i centra danych oraz stosowanie zaawansowanych technologii.

# Obszary zastosowań



Sales



Marketing



IT

# Obszary zastosowań



Sales



Marketing

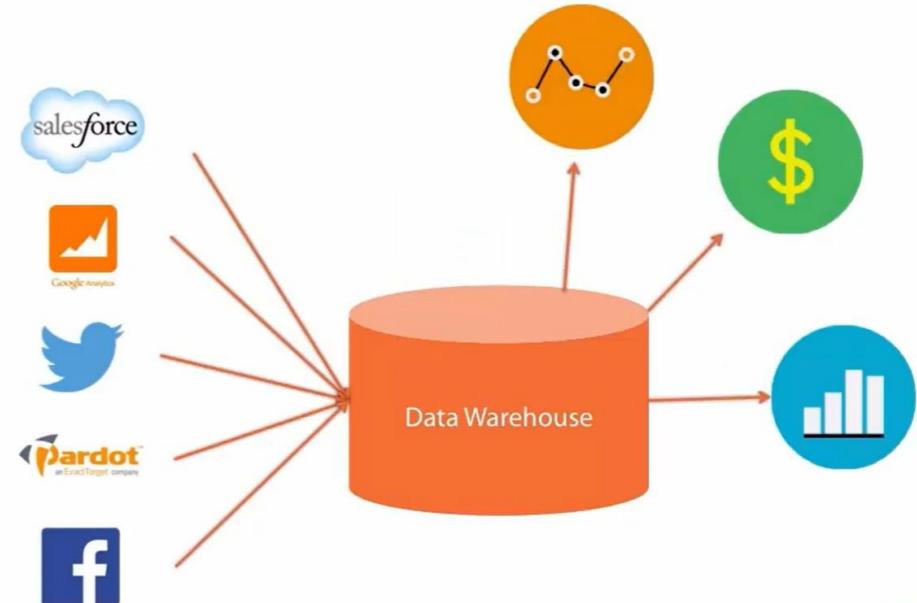
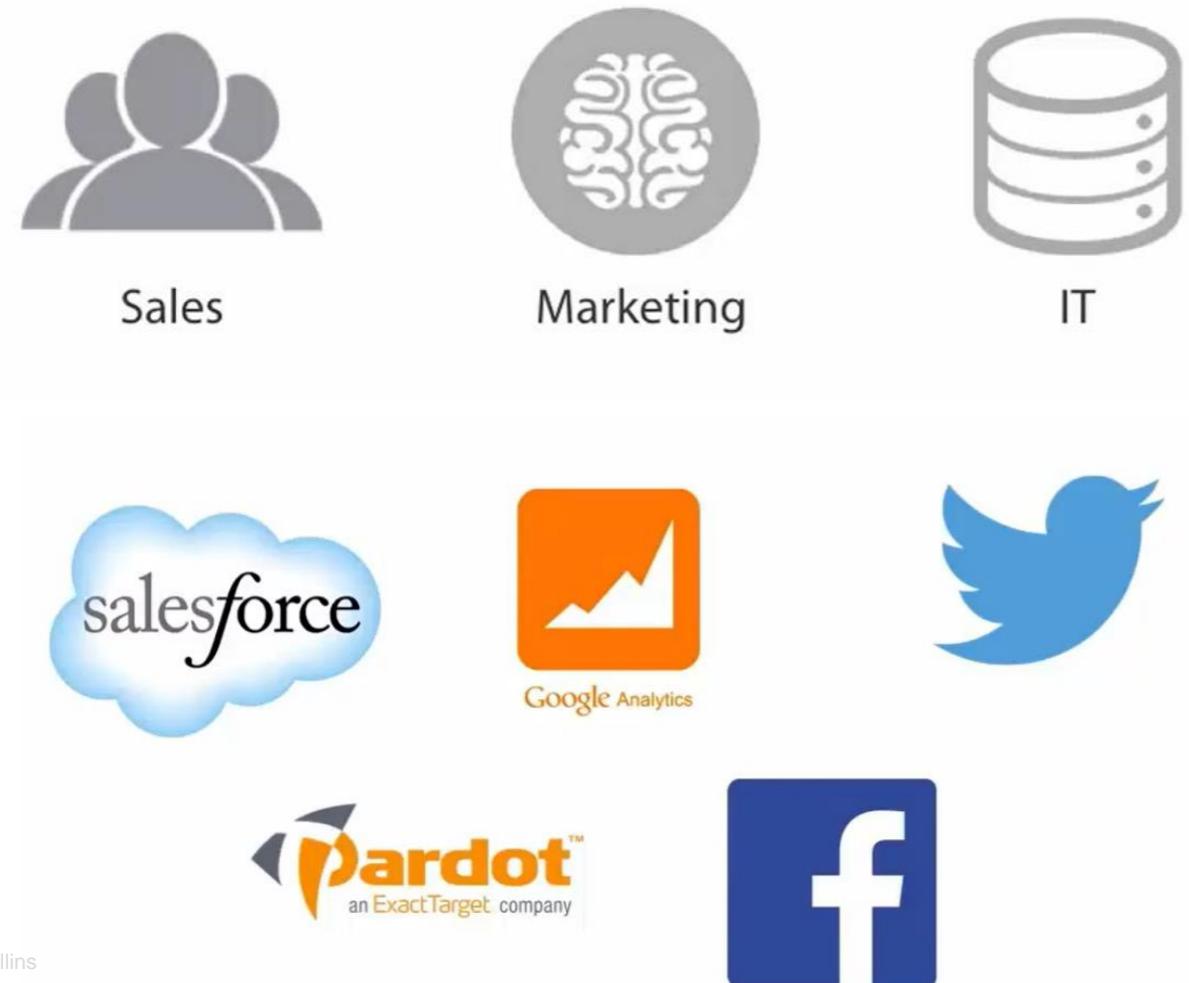


IT



Graphics: Ben Sullins

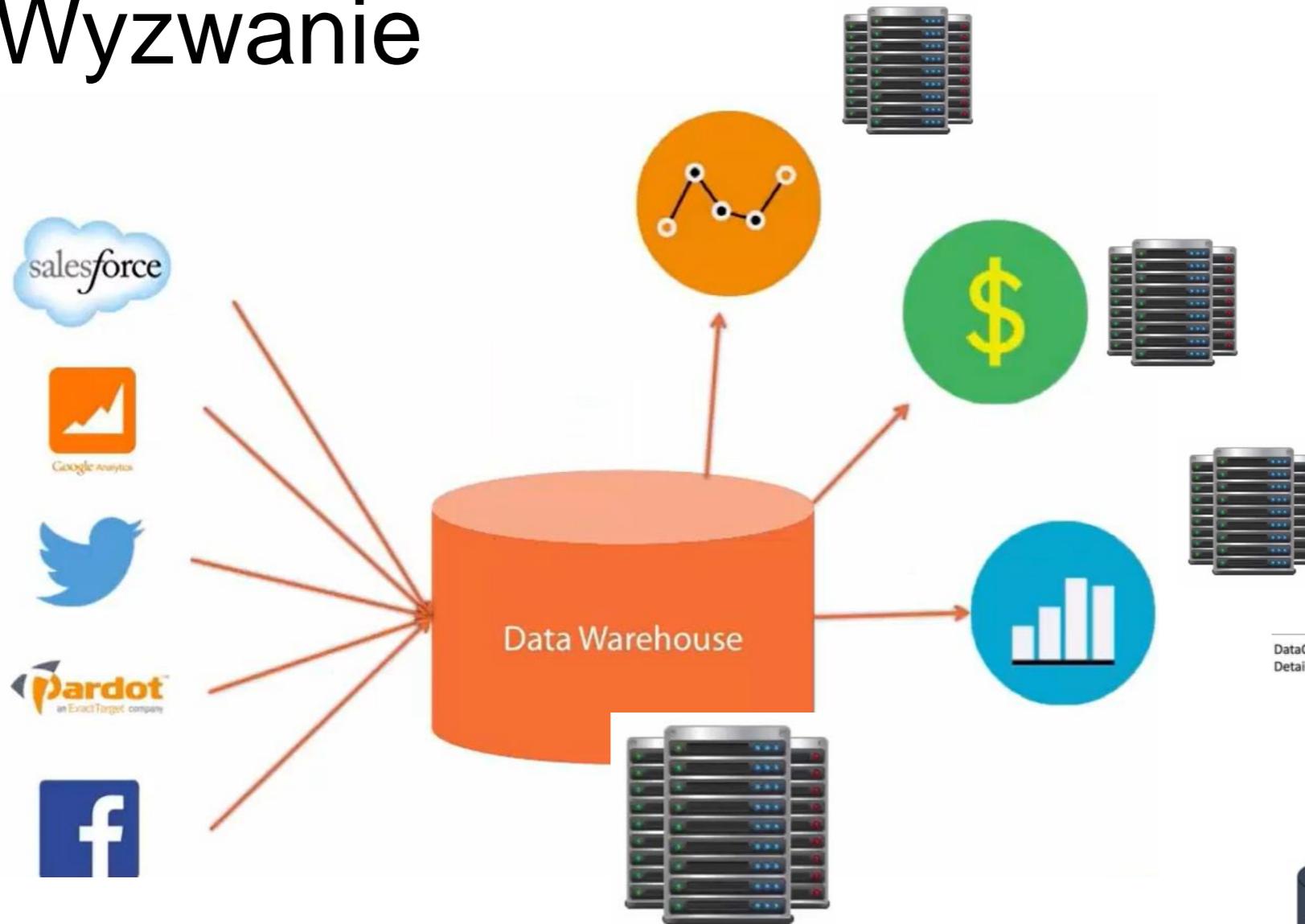
# Obszary zastosowań



Graphics: Ben Sullins

# Wyzwanie

Koszty, czas



Graphics: Ben Sullins

DataCone Big Data Cluster Detailed Infrastructure										Total Cpu: 128 Total Ram: 512 GB Total Disk: 40 TB	
Namenode	Cpu: 16 Ram: 64 GB Disk: 2 TB	Snamenode	Cpu: 8 Ram: 32 GB Disk: 2 TB	Kafka Broker	HDFS Namenode	HDFS Datanode	YARN	Spark2	Zookeeper	Airflow	
node01	Cpu: 16 Ram: 64 GB Disk: 2,5 TB	node02	Cpu: 8 Ram: 32 GB Disk: 2,5 TB	node03	Cpu: 8 Ram: 32 GB Disk: 2,5 TB	node04	Cpu: 16 Ram: 32 GB Disk: 1,5 TB	node05	Cpu: 16 Ram: 32 GB Disk: 1 TB	node06	Cpu: 8 Ram: 32 GB Disk: 2 TB
HDFS Datanode	YARN	Spark2	HDFS Datanode	YARN	Spark2	HDFS Datanode	YARN	Spark2	HDFS Datanode	YARN	
node07	Cpu: 16 Ram: 32 GB Disk: 2 TB	node08	Cpu: 8 Ram: 32 GB Disk: 2 TB	node09	Cpu: 8 Ram: 32 GB Disk: 1 TB					Airflow	

# Wyzwanie

2

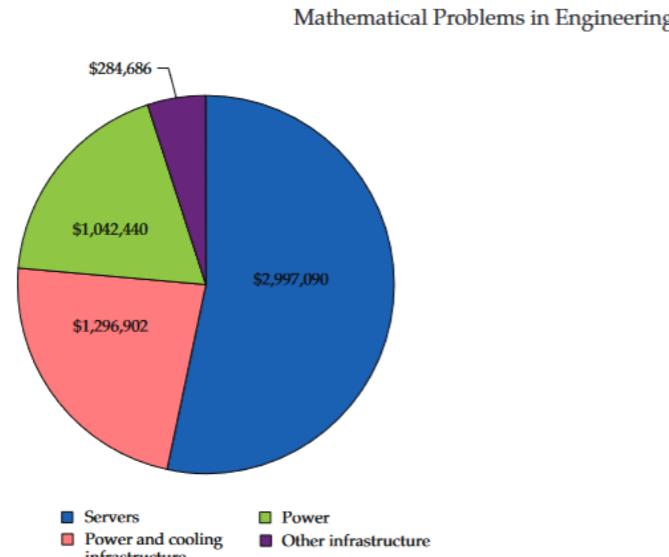


Figure 1: Monthly costs of the data center.

for energy consumption of all these facilities. In order to illustrate the importance of energy consumption for data centers, we introduce the concept, power usage effectiveness (PUE), which was developed by a consortium called The Green Grid.

**Definition 1.1.** Power usage effectiveness [3] is the ratio of total amount of power used by a data center facility to the power delivered to computing equipment. It is a measure of how efficiently a computer data center uses its power:

$$\text{PUE} = \frac{\text{Total facility power}}{\text{IT equipment power}}. \quad (1.1)$$

## *Power Usage Effectiveness (PUE)* *Carbon Usage Effectiveness (CUE)* *Water Usage Effectiveness (WUE)*

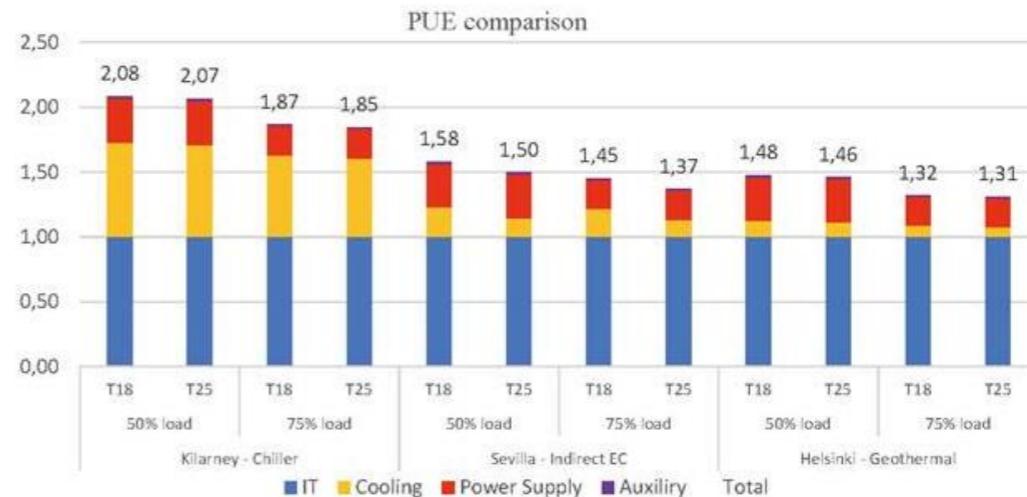


Figure 1. PUE results for the twelve different simulations for the three chosen configurations.

- *Energy-Efficient Multi-Job Scheduling Model for Cl.pdf*  
[https://www.researchgate.net/publication/258385511\\_Energy-Efficient\\_Multi-Job\\_Scheduling\\_Model\\_for\\_Cloud\\_Computing\\_and\\_Its\\_Genetic\\_Algorithm/link/56aa614608aed5a01358973b/download](https://www.researchgate.net/publication/258385511_Energy-Efficient_Multi-Job_Scheduling_Model_for_Cloud_Computing_and_Its_Genetic_Algorithm/link/56aa614608aed5a01358973b/download)
- *https://www.rehva.eu/rehva-journal/chapter/analysis-of-performance-metrics-for-data-center-efficiency-should-the-power-utilization-effectiveness-pue-still-be-used-as-the-main-indicator-part-2*

# Wyzwanie

Parameter	Description
Carbon usage Effectiveness(CUE)	It is a measure of carbon dioxide emission in environment by the data center. CUE=ECO2/EI Where , ECO2= Total carbon dioxide emission from total energy absorbed by the facility of a data center. EI = Total energy consumed by IT equipments.
Water Usage Effectiveness(WUE)	It is a measure of required water by a data center annually. Its defined as: WUE=Water used annually/EI
Data Center Productivity(DCP)	It is a measure of amount of fruitful work yielded by datacenter. It is defined as- DCP=Useful Work-done/Resource where, Resource = total resource taken to produce this useful work
Thermal Design Power(TDP)	It is the measurement of maximum amount of power required by cooling of computer system to dissipate. It is the maximum amount of power which a computer chip can take when running a real application.
Power Usage Effectiveness(PUE)	It is used for comparison of energy used by computing application and infrastructure equipment and the energy wasted in overhead. PUE = Total Facility Power/IT Equipment Power
Data Center Infrastructure Efficiency(DCIE)	It is the reciprocal of PUE. PUE and DCIE are most commonly used metrics that were designed for the comparison of efficiency of datacenters. It is defined as: DCIE=1/PUE DCIE = IT Equipment Power/Total Facility Power
Performance per Watt	It is the processing rate that can be remitted by a processor for each watt of power absorbed by it. This must be high. It measures the rate of computation that can be delivered by a computer for every watt of power consumed by it.
Green Energy Coefficient(GEC)	It is a measure of green energy (energy that comes from renewable sources) that is used by the facility of a datacenter. Energy consumed is measured in kWh. It is defined as GEC=Green Energy Consumed/Total Energy Consumed
Compute Power Efficiency(CPE)	It is a measure of the computing efficiency of a datacenter. As each watt consumed by server or cluster did not draw fruitful work all the time, some facility consumed power even in idle state and some consumes power for computing. CPE is defined as- CPE=IT Equipment Utilization/PUE =(IT Equipment Utilization*IT Equipment Power)/Total Facility Power
Energy reuse factor(ERF)	It is a measure of reusable energy (energy that is reused outside of a datacenter) that is used by datacenter. ERF=Re-used Energy Used/Total Energy Consumed

Table 1  
PARAMETERS USED FOR POWER PERFORMANCE

International Journal of Trend in Research and Development Volume -1(1)

## Green Cloud Computing: An Overview

Archana Gondalia  
M.Tech Network Technology  
Department of Computer Science and Engineering  
Institute Of Technology, Nirma University  
Ahmedabad, India

Neema Vyas  
M.Tech Network Technology  
Department of Computer Science and Engineering  
Institute Of Technology, Nirma University  
Ahmedabad, India

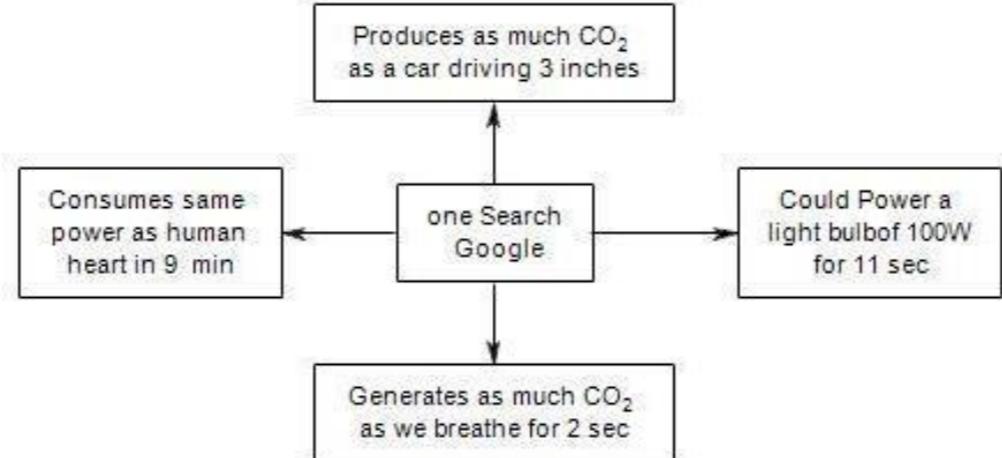


Figure 1. Energy utilised in one google search

The expansion of cloud computing technology, which provides computing resources over the internet, along with the explosive development of data-intensive technologies, such as big data and artificial intelligence, has increased the demand for high-density data centers.

**According to the International Energy Agency (IEA), the estimated global electricity consumption by data centers in 2022 was between 240 and 340 TWh [1], accounting for approximately 1–1.5% of the total global electricity consumption across all sectors. [2,3,4].** Given this considerable energy consumption, data centers are actively pursuing strategies to improve their energy efficiency [5,6].

Kim, Ji Hye, Dae Uk Shin, and Heegang Kim. "Data Center Energy Evaluation Tool Development and Analysis of Power Usage Effectiveness with Different Economizer Types in Various Climate Zones." *Buildings* 14.1 (2024): 299.

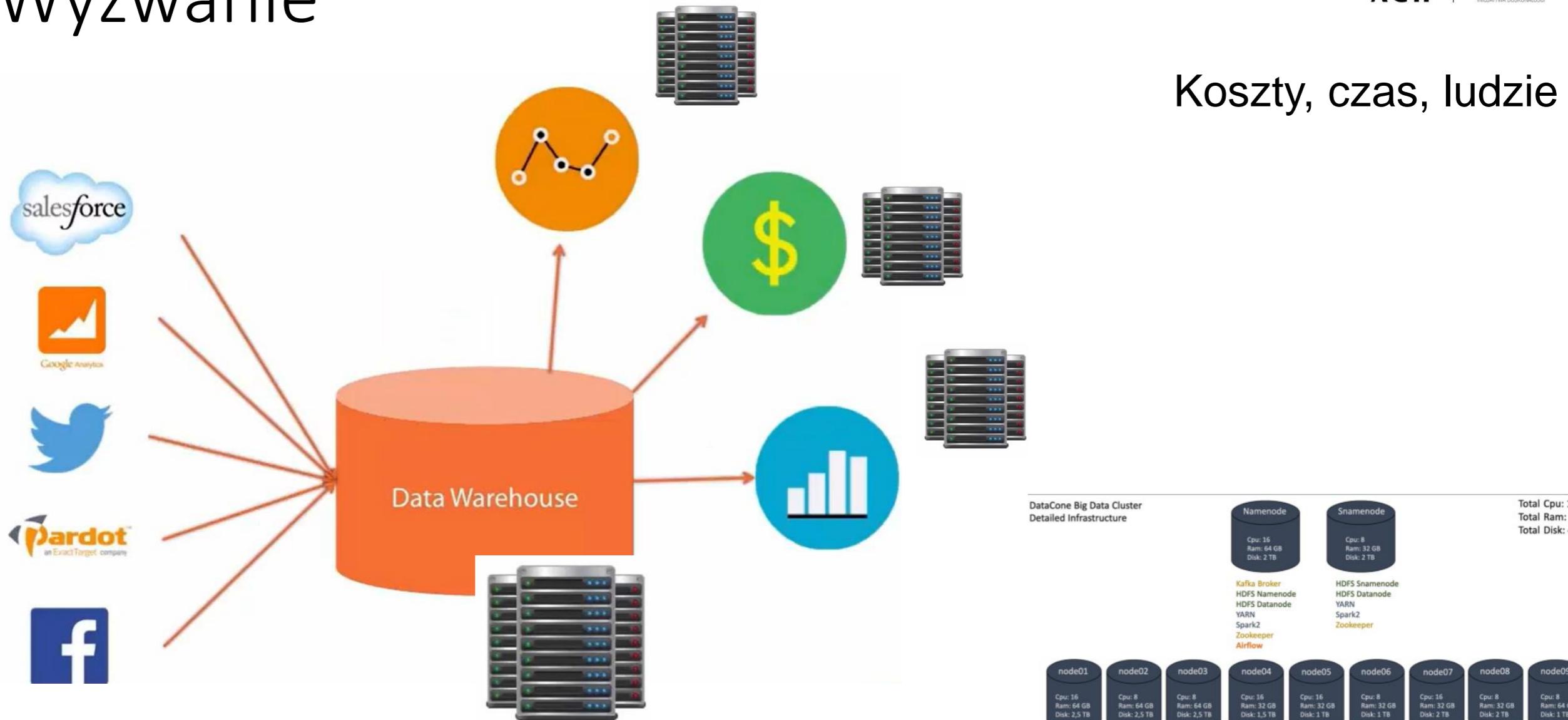
# Wyzwanie

Datacenter	Space	Energy	Sustainability	Property	Highlight
Harbin Data center	2.000.000 m <sup>2</sup>	150 MW	Unknown	China Mobile	Building design & gardens
Kolos Datacenter	6.500.000 ft <sup>2</sup> 600.000 m <sup>2</sup>	70 MW (up to 1000MW)	Will be 100% renewable energy	Kolos	Perfectly integrated with the natural landscape
The Citadel	1.300.000 ft <sup>2</sup> 120.037 m <sup>2</sup> (up to 17.4M ft <sup>2</sup> )	650 MW (fully operative)	100% renewable sources	Switch	260 patented innovations (construction & operation)
China Telecom Inner Mongolia Inform. Park	3.200.000m <sup>2</sup> 295.475 m <sup>2</sup>	150 MW	Unknown	China Telecom	Colocation Facility Cloud Node
Utah Datacenter	5.415.000 ft <sup>2</sup> 500.000 m <sup>2</sup>	65 MW	Unknown	Databank	Cloud Node Colocation Facility DRBC Site

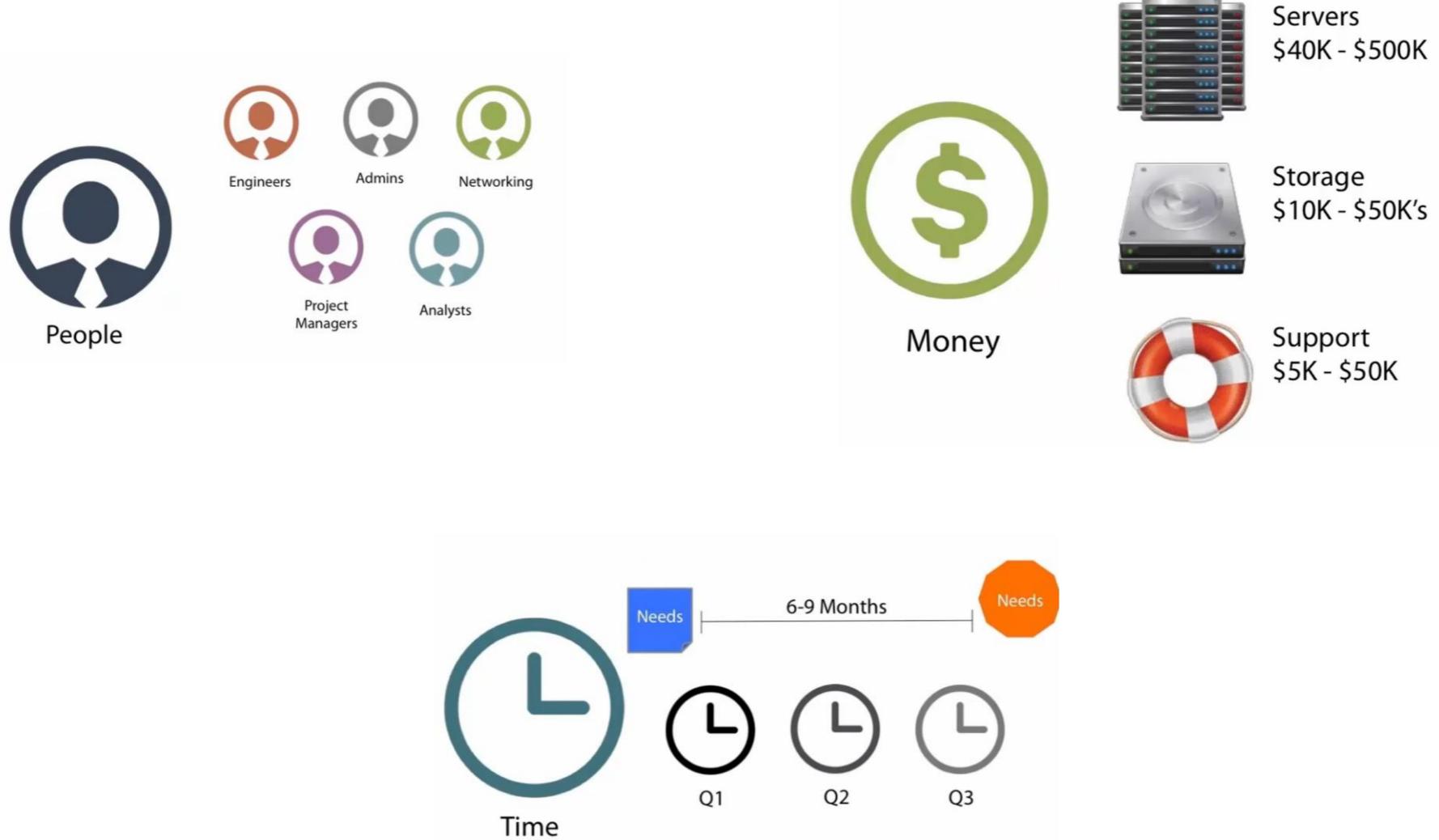
Założenia/Plany:

UE -> Neutralnie emisyjne ICT do 2030 ?

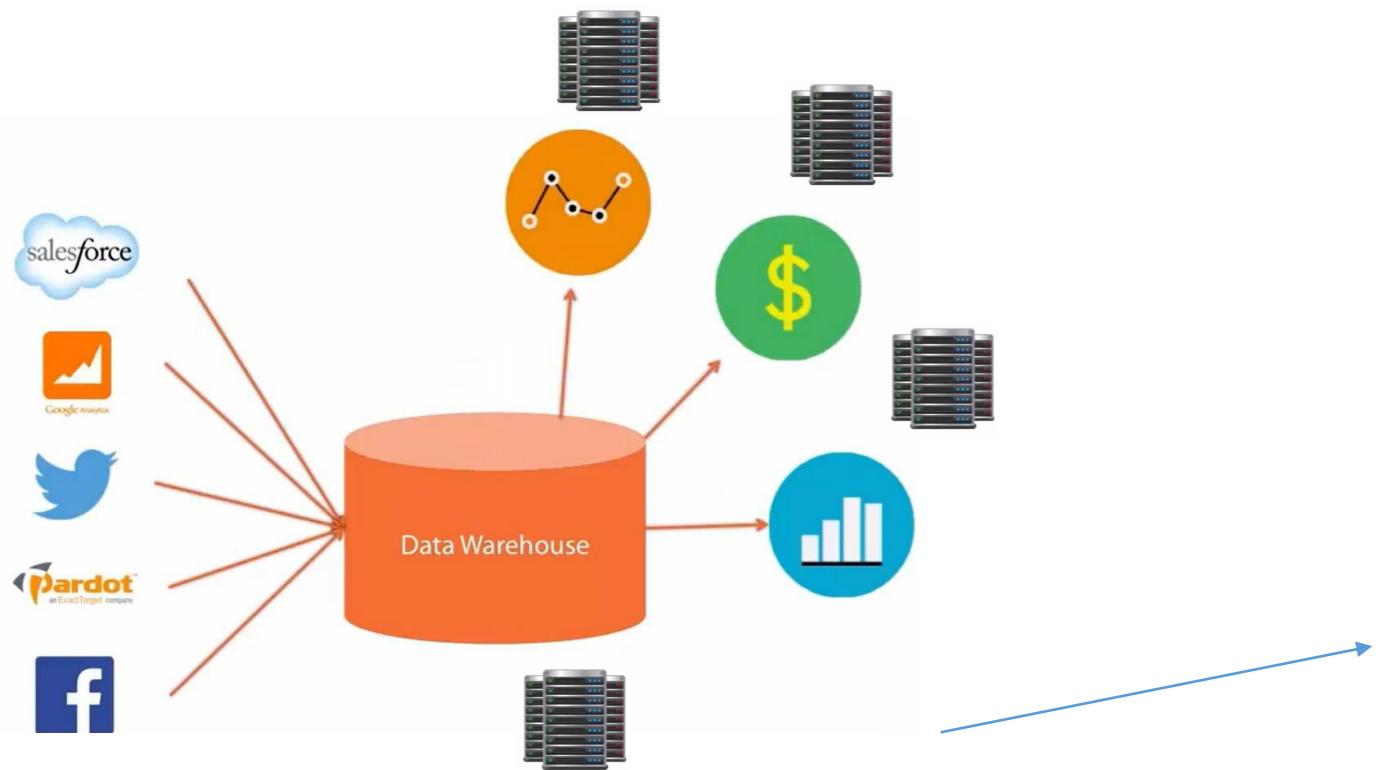
# Wyzwanie



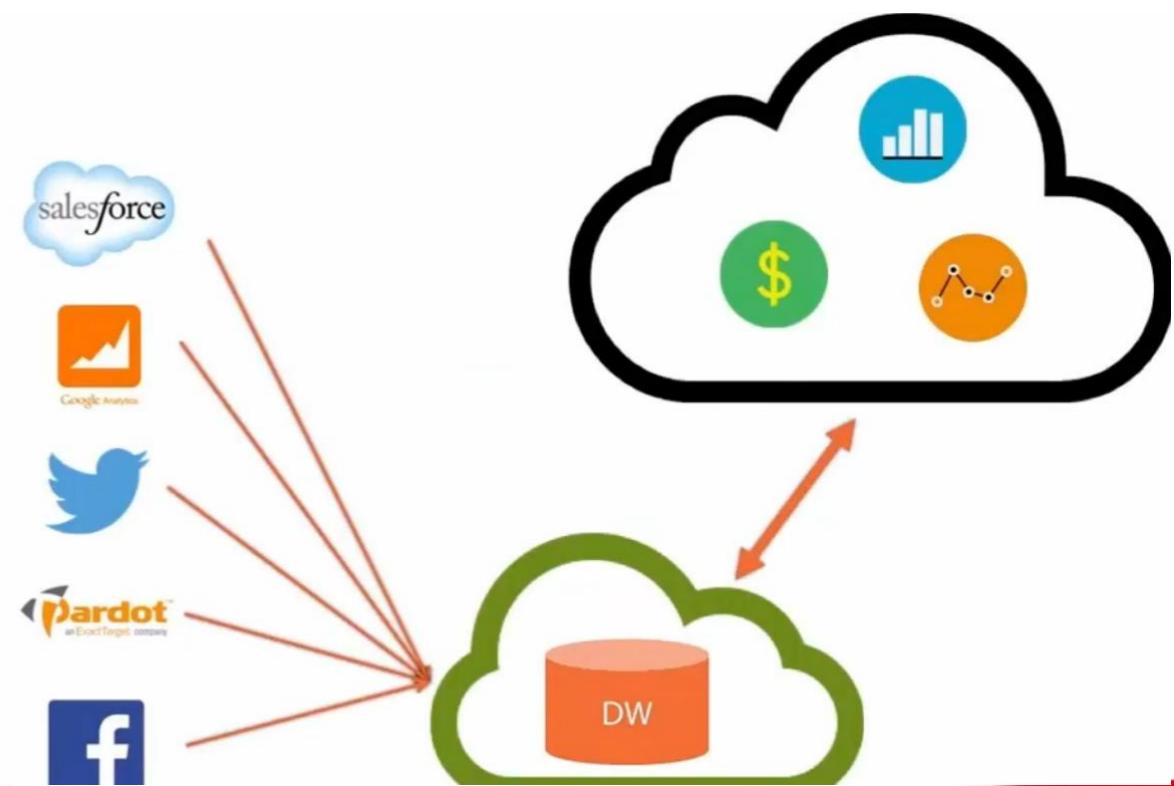
# Wyzwanie



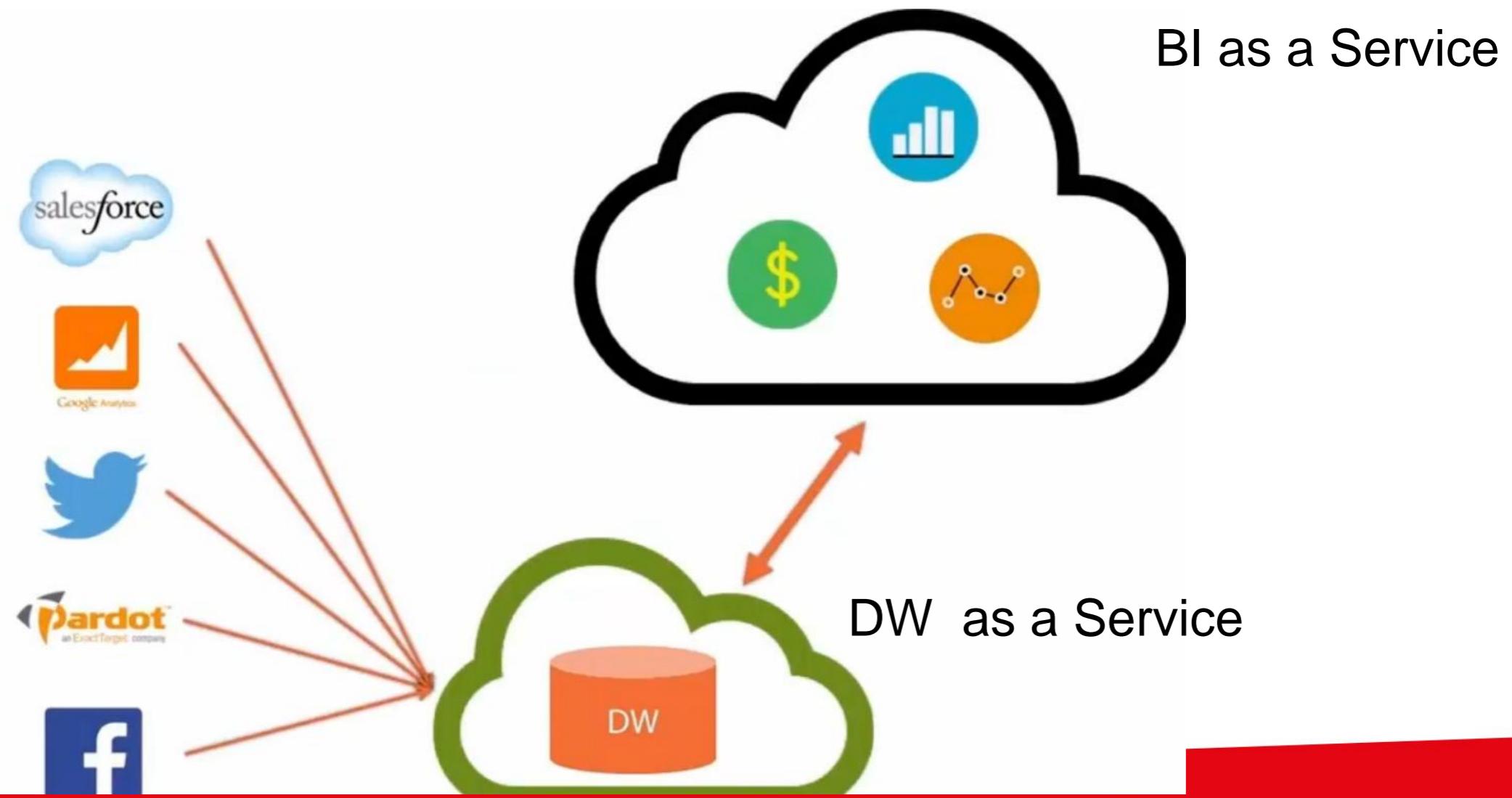
# On-Premise vs Cloud



	On-Premise	Cloud
5 FTE	1 FTE	
3 Months	2 Weeks	
\$150K+ /yr	\$5-6K/yr	



# BI as a Service DW as a Service



# BI as a Service

## Basic Charting

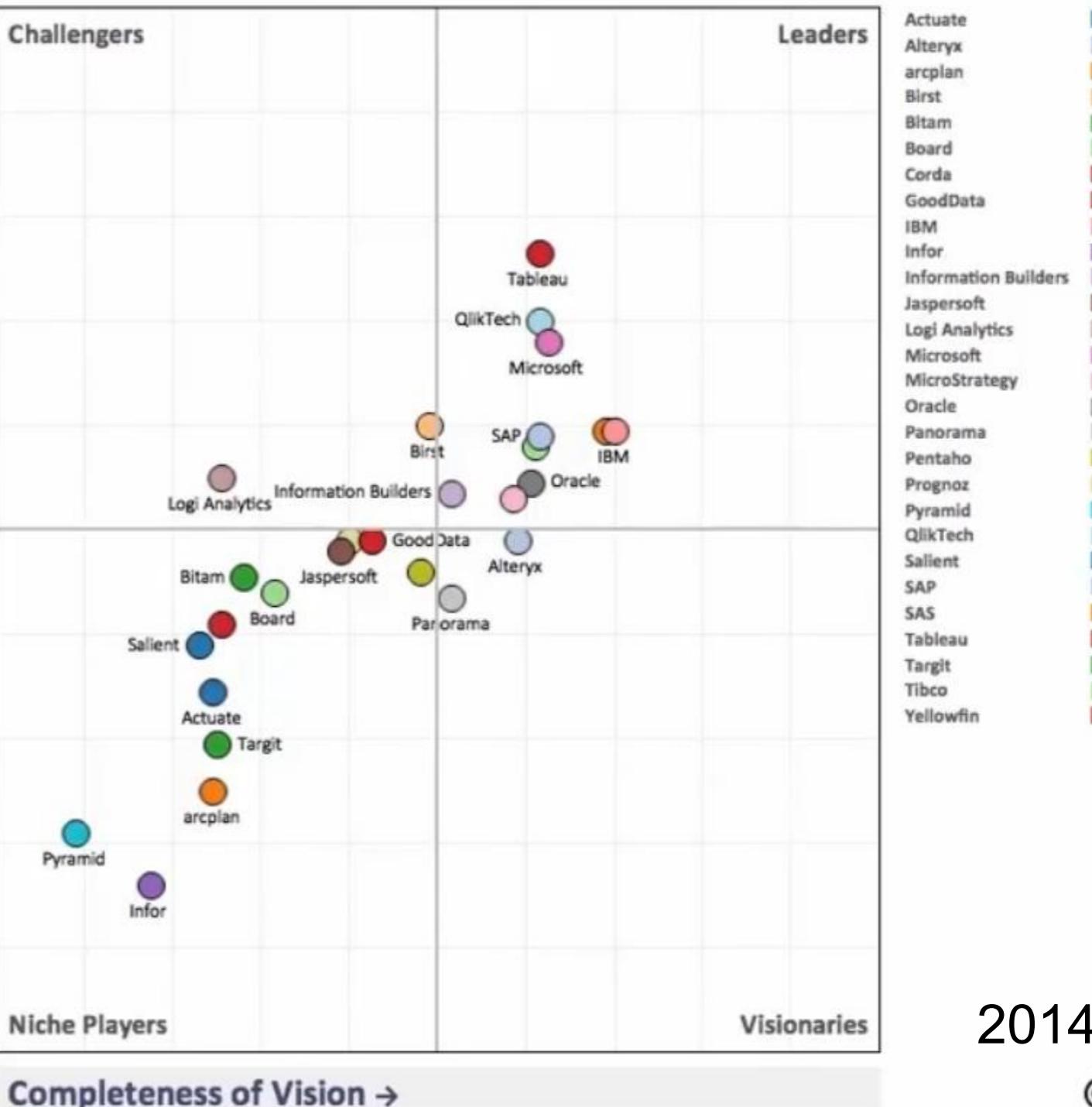


Przykładowe wyzwania – wydajność czytania danych z HTTPs etc.  
Dużo rozwiązań na rynku (mniej lub bardziej kompleksowych)

# BI as a Service

## Kompleksowość

Gartner Magic Quadrant for Business Intelligence and Analytics Platforms



<https://www.gartner.com/reviews/market/analytics-business-intelligence-platforms>

<https://www.gartner.com/en/documents/4247699?ref=null>  
agn.edu.pl

2014 ..



# BI as a Service

Gartner Magic Quadrant for Business Intelligence and Analytics Platforms

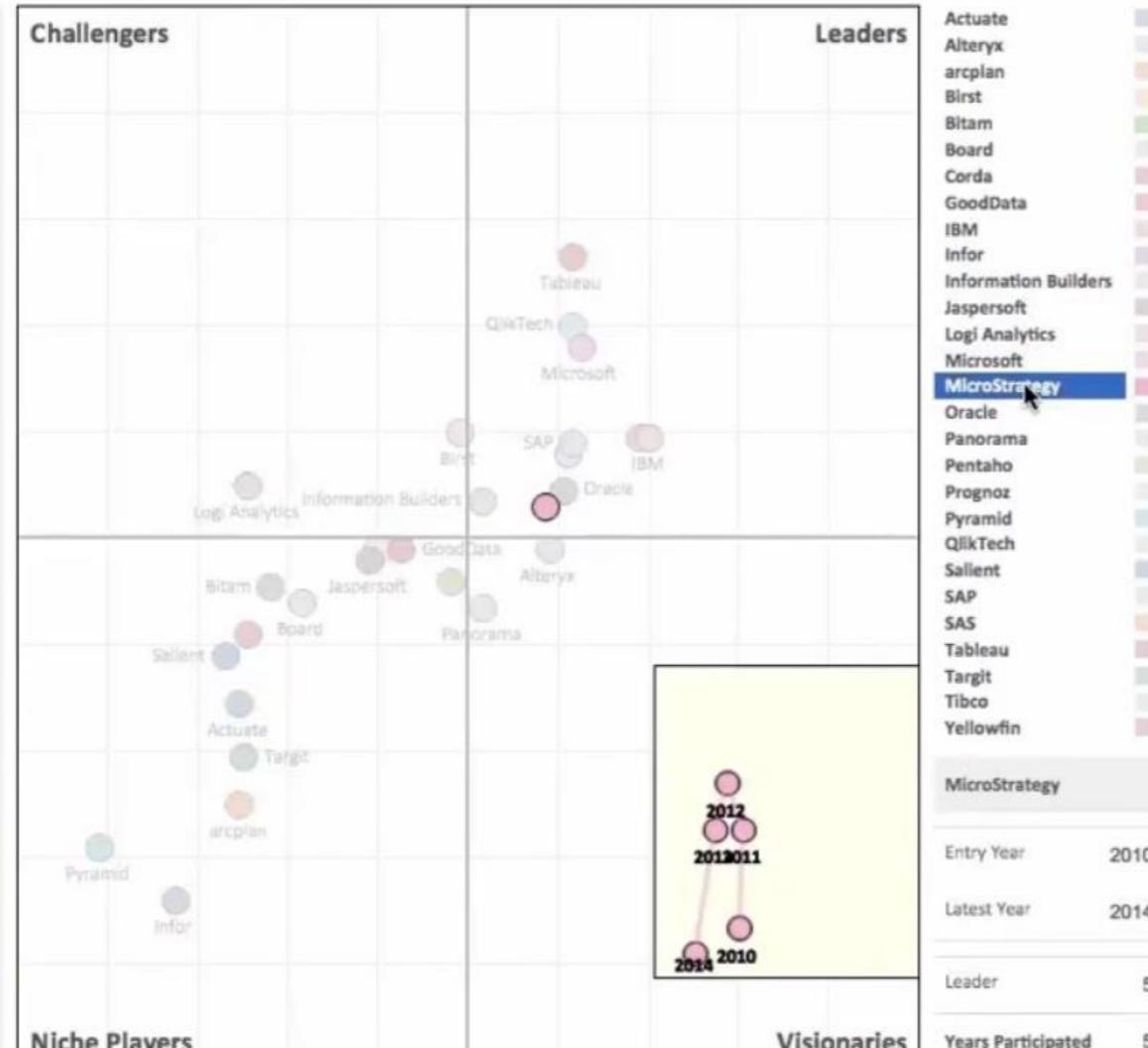


Gartner Magic Quadrant for Business Intelligence and Analytics Platforms

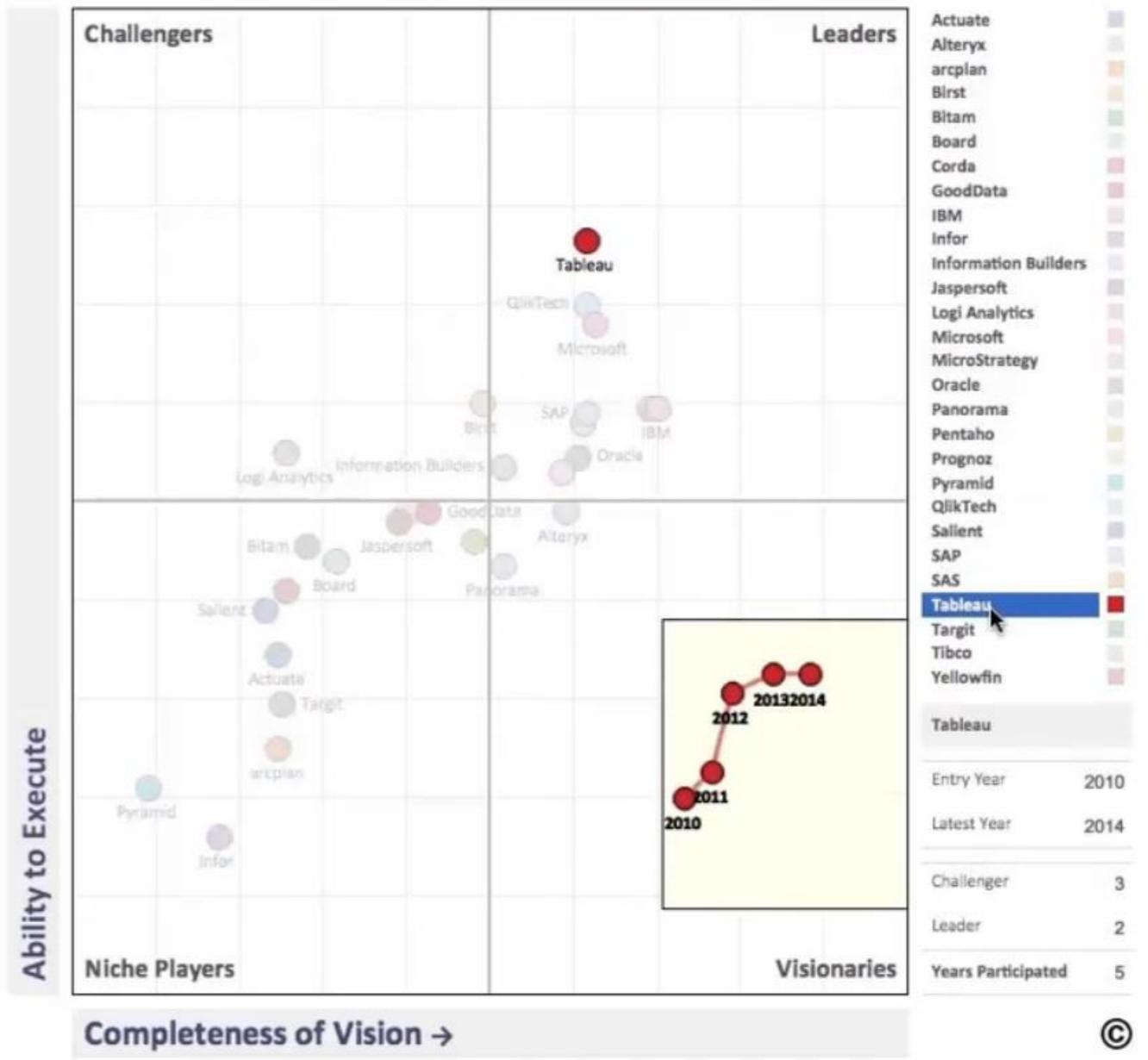


# BI as a Service

Gartner Magic Quadrant for Business Intelligence and Analytics Platforms



Gartner Magic Quadrant for Business Intelligence and Analytics Platforms



Completeness of Vision →

©

# DW as a Service

Kompleksowość



Amazon Redshift



Google BigQuery

Hosted, rozproszona DB

- Np. Różnią się, ponieważ niektórzy Vendorzy (np. Microsoft) oferuje Infrastrukturę w chmurze.
- Koszty utrzymania – ile mamy danych itd.



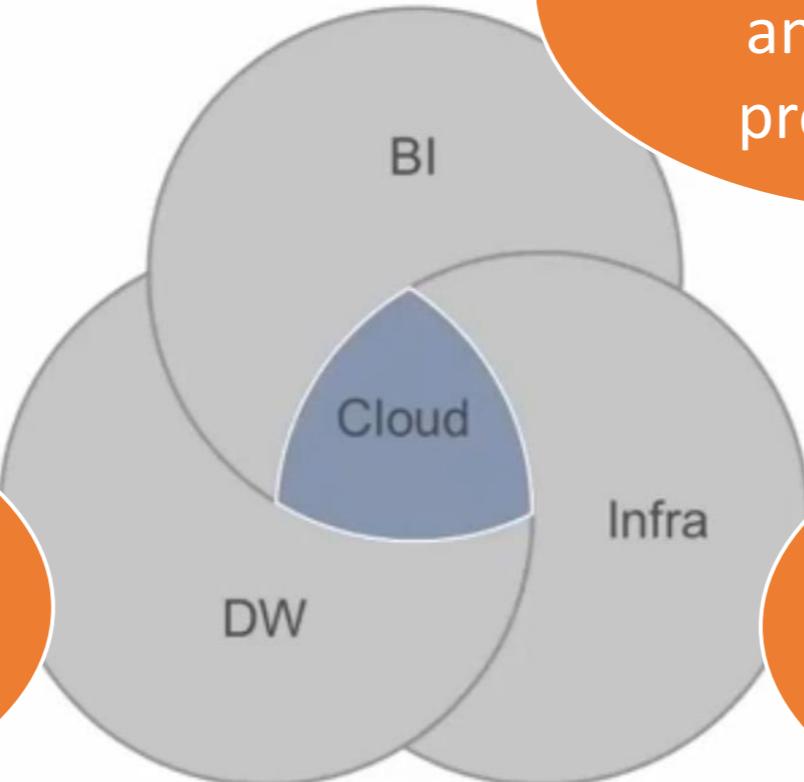
Amazon EC2



Google Compute Engine

# Usługi

Przetrzymywanie danych (storage), dostęp do danych (query access)

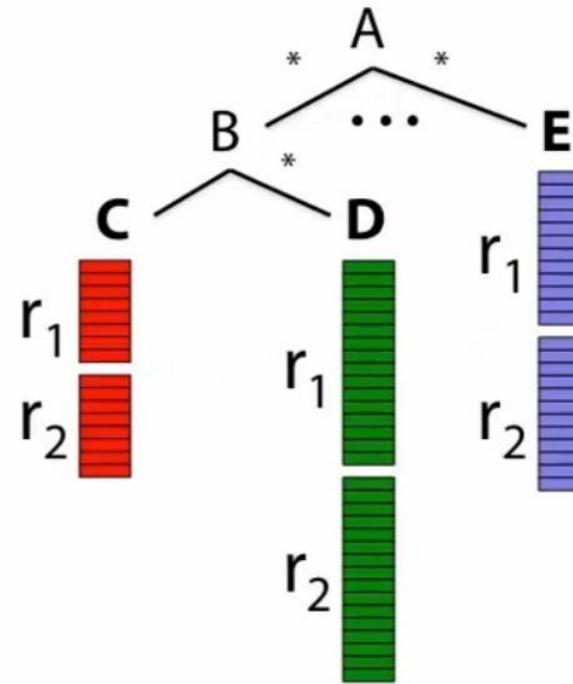
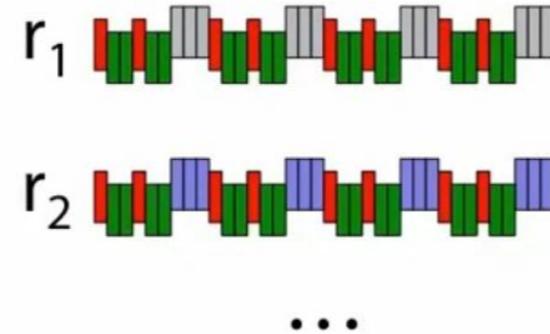


Data acquisition,  
cleansing,  
analysis and  
presentation

Wirtualne  
serwery (virtual  
servers), hostingi

# Przykład (Google Big Query)

Magazynowa  
nie danych –  
kolumny vs  
wiersze –  
skanowanie  
wszystkich  
rekordów  
Magazynujem  
y zestawy  
kolumn.



Row Storage

1:Office,Central,6  
2:Office,West,13  
3:Technology,West,50  
4:Furniture,West,50  
5:Furniture,South,20  
6:Technology,South,24  
7:Technology,West,20

Column Storage

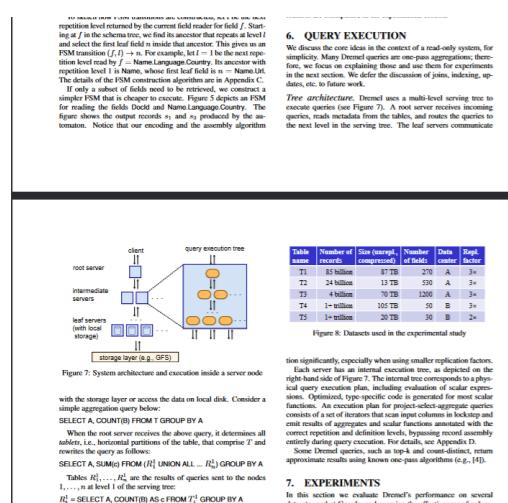
{Office:1,2},{Technology:3,6,7},{Furniture:4,5}  
{Central:1},{West:2,3,4,7},{South:5,6}  
{6:1},{13:2},{50:3,4},{20:5,7},{24:6}

<http://static.googleusercontent.com/media/research.google.com/en/us/pubs/archive/36632.pdf> (Dremel: Interactive Analysis of Web-Scale Datasets  
 Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer,  
 Shiva Shivakumar, Matt Tolton, Theo Vassilakis  
 Google, Inc)

# Przykład (Google Big Query)

Zapytania - architektura rozproszzonego systemu przetwarzania zapytań za pomocą wielopoziomowego drzewa

Art.: Dremel – poprzednik BQ



„System przetwarza zapytania w równoległej i rozproszonej architekturze drzewa, gdzie klient wysyła zapytanie do głównego serwera, który przekazuje je do serwerów pośrednich. Te tworzą mniejsze plany wykonania zapytania i przesyłają je do serwerów liści, które wykonują faktyczne operacje na danych, a wyniki są agregowane i przesyłane z powrotem do klienta, co umożliwia bardzo szybkie przetwarzanie na dużą skalę”

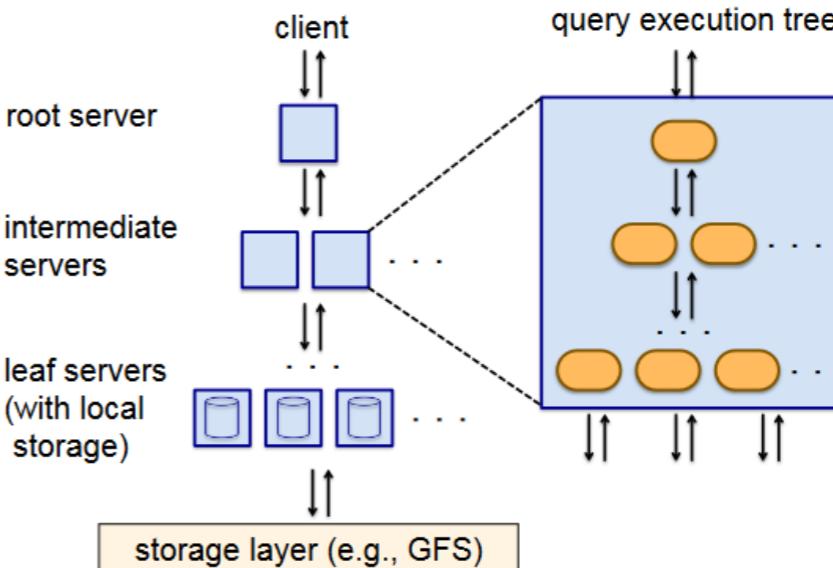


Figure 7: System architecture and execution inside a server node

# Przykład (Google Big Query)

- Analysis of crawled web documents.
- Tracking install data for applications on Android Market.
- Crash reporting for Google products.
- OCR results from Google Books.
- Spam analysis.
- Debugging of map tiles on Google Maps.
- Tablet migrations in managed Bigtable instances.
- Results of tests run on Google's distributed build system.
- Disk I/O statistics for hundreds of thousands of disks.
- Resource monitoring for jobs run in Google's data centers.
- Symbols and dependencies in Google's codebase.

# Przykład (Google Big Query)



The screenshot shows the Google Cloud BigQuery Pricing page. The top navigation bar includes links for Overview, Solutions, Products, Pricing (which is highlighted), and Resources. The main content area features a search bar, navigation links for Docs, Support, Language, Console, and user profile, and buttons for Contact Us and Start free. On the left, a sidebar titled 'PRICING' lists various pricing categories: BigQuery pricing, Overview of BigQuery pricing, On-demand compute pricing, Capacity compute pricing, Storage pricing, Data Transfer Service pricing, BigQuery Omni pricing, Data ingestion pricing, Data extraction pricing, Data replication pricing, External Services, BigQuery ML pricing, BI Engine pricing, Free operations, Free usage tier, Flat-rate pricing, and What's next.

## BigQuery pricing

BigQuery is a serverless data analytics platform. You don't need to provision individual instances or virtual machines to use BigQuery. Instead, BigQuery automatically allocates computing resources as you need them. You can also reserve compute capacity ahead of time in the form of slots, which represent virtual CPUs. The pricing structure of BigQuery reflects this design.

### Overview of BigQuery pricing

BigQuery pricing has two main components:

- [Compute pricing](#) is the cost to process queries, including SQL queries, user-defined functions, scripts, and certain data manipulation language (DML) and data definition language (DDL) statements.
- [Storage pricing](#) is the cost to store data that you load into BigQuery.

BigQuery charges for other operations, including using [BigQuery Omni](#), [BigQuery ML](#), [BI Engine](#), and streaming [reads](#) and [writes](#).

In addition, BigQuery has [free operations](#) and a [free usage tier](#).

Every project that you create has a billing account attached to it. Any charges incurred by BigQuery jobs run in the project are billed to the attached billing account. BigQuery storage charges are also billed to the attached billing account. You can view BigQuery costs and trends by using the Cloud Billing reports page in the Google Cloud console.

 **Key Point:** Pricing models apply to accounts, not individual projects, unless otherwise specified.

#### Compute pricing models

BigQuery offers a choice of two compute pricing models for running queries:

- *Odwiedź oficjalną stronę cennika*
- *Przejrzyj szczegóły cen*
- *Zrozum limity i zakres bezpłatnych operacji*
- *Sprawdź aktualizacje i zmiany*
- *Korzystaj z kalkulatorów danych (strony vendorów)*
- *Pamiętaj o możliwościach ładowania danych:*

#### Strumieniowe vs Wsadowe:

**Strumieniowe:** Dane są przesyłane bezpośrednio do w czasie rzeczywistym. Jest to szczególnie przydatne w przypadkach, kiedy dane muszą być natychmiast dostępne do analizy.

**Wsadowe:** Dane są przesyłane w dużych partiami, często według ustalonego harmonogramu, na przykład raz na dobę. Jest to bardziej kosztowo efektywne dla dużych ilości danych, które nie wymagają natychmiastowej analizy.

# Przykład (Google Big Query Docs)

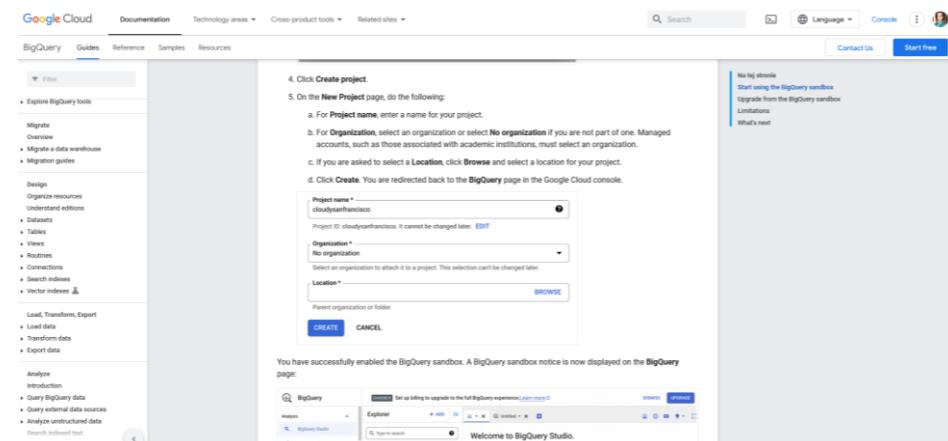
- Load, Transform, Export**
- ▶ Load data
  - Introduction
  - ▶ BigQuery Data Transfer Service
  - ▶ Batch load data
  - ▶ Write and read data with the Storage API
- Load data from other Google services
- Load data using third-party apps
- Load data using cross-cloud operations
- ▶ Transform data
- ▶ Export data

- Analyze**
- Introduction
- ▶ Query BigQuery data
- ▶ Query external data sources
- ▶ Analyze unstructured data
- Search indexed text
- Work with text analyzers
- ▶ Work with sessions
- ▶ Use geospatial analytics
- ▶ Use programmatic tools
- ▶ Use analysis and BI tools
- ▶ Share with Analytics Hub

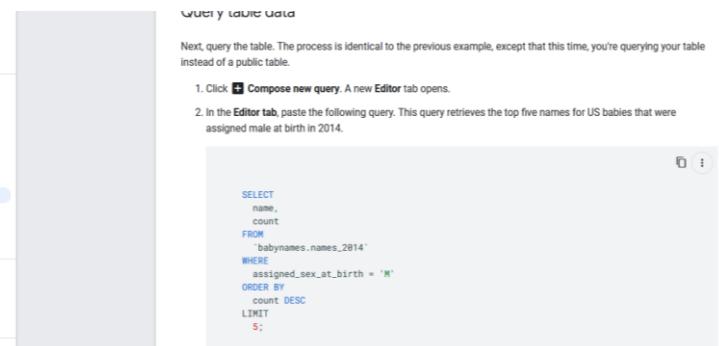
Big Query Setup (Console etc), MyProjects, APIs, Billing, WebQL

Example:

[https://cloud.google.com/bigquery/docs/sandbox?\\_gl=1\\*18xoz2d\\*\\_up\\*MQ..&gclid=CjwKCAjwi\\_exBhA8EiwA\\_kU1MgMLye9ki77Z7tFL041odOkpMTFyhR3hB1eQ4HW0CqUXWVR5EVfcDRoCFy4QAvD\\_BwE&gclsrc=aw.ds](https://cloud.google.com/bigquery/docs/sandbox?_gl=1*18xoz2d*_up*MQ..&gclid=CjwKCAjwi_exBhA8EiwA_kU1MgMLye9ki77Z7tFL041odOkpMTFyhR3hB1eQ4HW0CqUXWVR5EVfcDRoCFy4QAvD_BwE&gclsrc=aw.ds)



The screenshot illustrates the initial steps of setting up a BigQuery project. It shows the 'Create project' dialog where the user has entered 'cloudyconfiance' as the project name. The 'Organization' dropdown is set to 'No organization'. A note at the bottom of the dialog states: 'You have successfully enabled the BigQuery sandbox. A BigQuery sandbox notice is now displayed on the BigQuery page.' Below this, the 'BigQuery' interface is visible, showing a 'Welcome to BigQuery Studio' message.

The screenshot shows the 'Load and query data' section of the BigQuery documentation. It features a 'Compose new query' editor tab. Inside the editor, there is a SQL query:

```

SELECT
    name,
    count
FROM
    `babynames.names_2014`
WHERE
    assigned_sex_at_birth = 'M'
ORDER BY
    count DESC
LIMIT
5;
  
```

Example

The following Python client example loads CSV data from a Google Cloud Storage bucket and prints the results on the command line.

```

# Python example
# Loads the table from Google Cloud Storage and prints the table.
def loadTable(service, projectId, datasetId, targetTableId, sourceCSV):
    try:
        jobCollection = service.jobs()
        jobData = {
            'projectId': projectId,
            'configuration': {
                'load': {
                    'sourceUris': [sourceCSV],
                    'schema': {
                        'fields': [
                            {
                                'name': 'Name',
                                'type': 'STRING'
                            },
                            {
                                'name': 'Age',
                                'type': 'INTEGER'
                            }
                        ]
                    }
                }
            }
        }
    
```

- Load, Transform, Export**
- ▶ Load data
  - Introduction
  - ▶ BigQuery Data Transfer Service
  - ▶ Batch load data
  - ▶ Write and read data with the Storage API
- Read data with the Storage Read API
- Write data with the Storage Write API
- Load data from other Google services
- Load data using third-party apps
- Load data using cross-cloud operations
- ▶ Transform data
- ▶ Export data

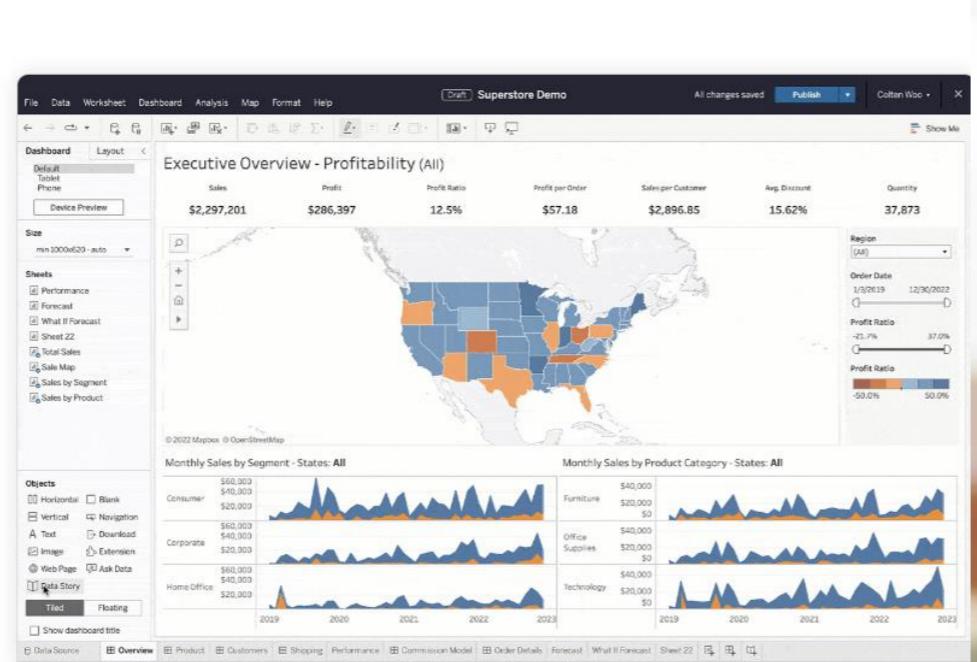
Example: <https://www.tableau.com/en-gb/products/cloud-bi>

## Tableau Cloud capabilities



### Drive faster, smarter decisions

Instil confident decision-making and harness the power of your data with intelligent tools like Data Stories, Ask Data and Explain Data. Save time and make analytics easy for everyone by adding [Data Stories](#), automated and easy-to-understand narratives, to dashboards. Explore and answer critical business questions in natural language with [Ask Data](#). With [Explain Data](#), discover the "why" behind AI-driven insights that invite deeper exploration. Map your data journey with [Tableau Blueprint](#) to become a data-driven organisation. Scale data-driven decision making with AI-powered insights from [Tableau Pulse](#) and integrate data into your day-to-day tasks.



**FAST ANALYTICS FOR EVERYONE**

**Tableau Desktop**

**BUSINESS INTELLIGENCE**

**Tableau Server**

**ANALYTICS IN THE CLOUD**

**Tableau Online**

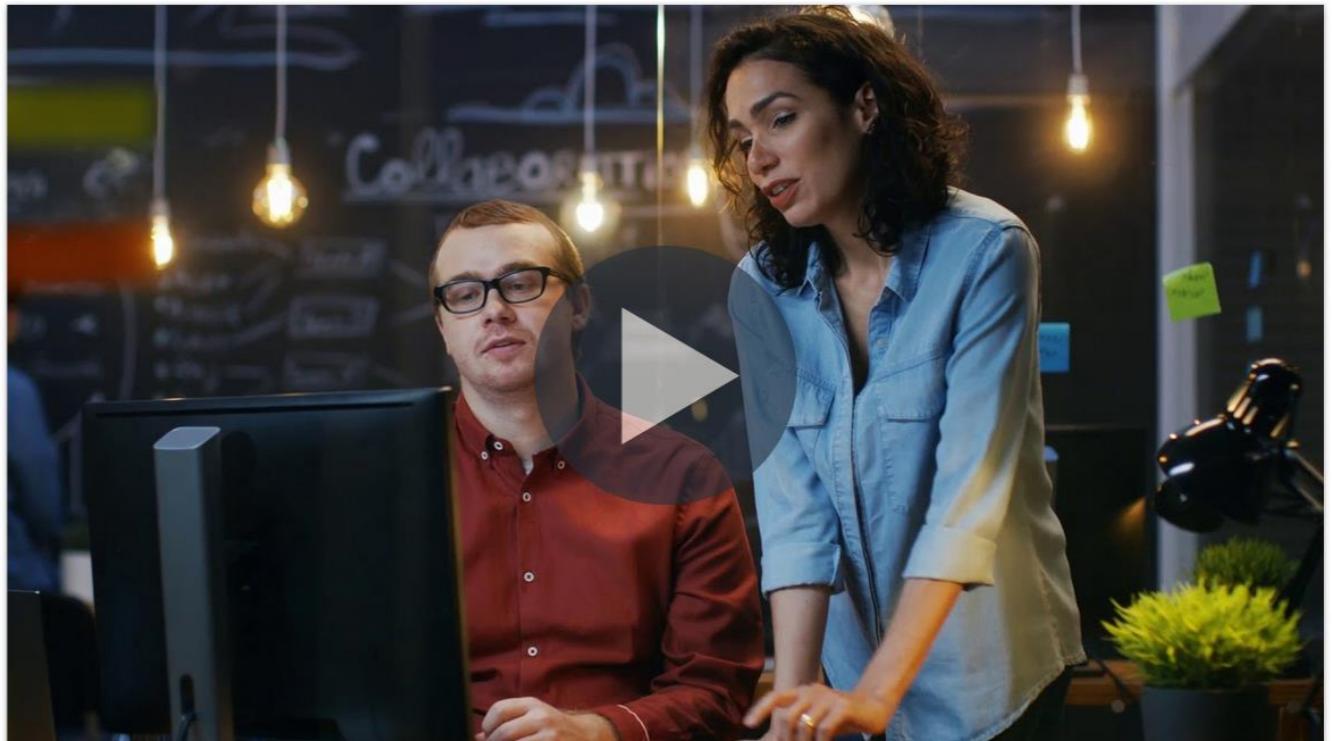


About AWS Contact Us Support▼ English▼ My Account▼ Sign In [Create an AWS Account](#)

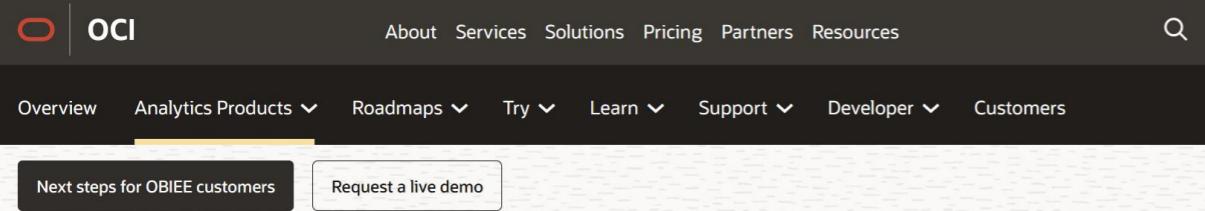
Products Solutions Pricing Documentation Learn Partner Network AWS Marketplace Customer Enablement Events Explore More 

## Why Amazon QuickSight

- ✓ Pay only for what you use
- ✓ Scale to tens of thousands of users
- ✓ Easily [embed analytics](#) to differentiate your applications
- ✓ Enable BI for everyone with [QuickSight Q](#)



Amazon QuickSight - Overview (2:01)



The screenshot shows the OCI homepage with a dark header bar. The top navigation includes links for About, Services, Solutions, Pricing, Partners, Resources, and a search icon. Below the header are secondary navigation links: Overview, Analytics Products, Roadmaps, Try, Learn, Support, Developer, and Customers. At the bottom of the page are two buttons: "Next steps for OBIEE customers" and "Request a live demo".

## Analytics Platform

The Oracle Analytics platform is a cloud native service that provides the capabilities required to address the entire analytics process including data ingestion and modeling, data preparation and enrichment, and visualization and collaboration, without compromising security and governance. Embedded machine learning and natural language processing technologies help increase productivity and build an analytics-driven culture in organizations. Start on-premises or in the cloud—Oracle Analytics supports a hybrid deployment strategy, providing flexible paths to the cloud.




## Explore Oracle Analytics

- Data visualization and storytelling**
- Machine learning
- Mobile analytics app
- Open data source connectivity
- Data preparation and enrichment
- Enterprise data modeling

### Build compelling visual stories using analytics

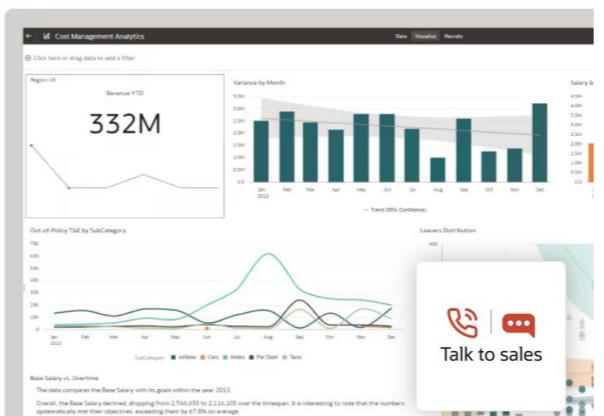
Visually explore data to create and share compelling stories using Oracle Analytics. Discover the signals in data that can turn complex relationships into engaging, meaningful, and easy-to-understand communications.

Accelerate the data analytics process and make decisions with actionable information. A code-free, drag-and-drop interface enables anyone in the organization to build interactive [data visualizations](#) without specialized skills.

[Learn more about the capabilities](#)

[See analytics in action \(3:07\)](#)

[Talk to sales](#)



Każda z tych platform posiada swoje unikalne mocne strony:

- **Google BigQuery** doskonale radzi sobie z efektywnym przetwarzaniem ogromnych zbiorów danych.
- **Microsoft Power BI** jest świetnym rozwiązaniem dla organizacji już zintegrowanych z ekosystemem Microsoft.
- **AWS QuickSight** oferuje elastyczne środowisko bezserwerowe, idealne dla startupów i średnich firm.
- **Tableau Online** jest preferowany ze względu na swoje potężne możliwości wizualizacji danych.
- **Oracle Analytics Cloud** i **SAP Analytics Cloud** są dobrze dopasowane do potrzeb przedsiębiorstw wymagających kompleksowego zestawu narzędzi obejmujących wszystko od przetwarzania danych po zaawansowaną analizę i planowanie.
- Itp..

Wybór odpowiedniego rozwiązania BI w chmurze w głównej mierze zależy od konkretnych potrzeb biznesowych, wielkości i charakteru danych oraz istniejącej infrastruktury technologicznej organizacji.

# Ćwiczenie 12

Proszę przygotować porównanie rozwiązań BI, które można wdrożyć w przykładowej organizacji. Porównanie powinno opierać się na kryteriach takich jak funkcjonalność na każdym etapie cyklu życia danych (DLC), koszty, architektura (on-premise/cloud itp.) oraz wsparcie. Porównanie powinno jasno ukazywać funkcjonalne możliwości różnych rozwiązań. Proszę przesyłać zadanie prowadzącemu do środy, **29.05.2024**.

**Przegląd rynku rozwiązań BI** skupiając się na co najmniej 4 różnych rozwiązaniach. Analiza powinna obejmować przynajmniej następujące aspekty każdego narzędzia:

- **Cena:** Koszt zakupu/subskrypcji, utrzymania, całosciowy wdrożenia (miesięcznie, rocznie).
- **Integracja i magazynowanie danych:** jakie są możliwości pozyskiwania/integracji i magazynowania danych;
- **Analiza i raportowanie:** Funkcjonalności związane z analizą danych, wizualizacją oraz generowaniem raportów.
- **Skalowalność:** Możliwość dostosowania narzędzia do rosnących potrzeb organizacji.
- **Wsparcie i społeczność:** Dostępność pomocy technicznej i zasobów społecznościowych.

## Macierz decyzyjna:

- **Tworzenie macierzy:** Utworzenie macierzy decyzyjnej, w której ocenione i porównane zostaną rozwiązania BI w odniesieniu do wcześniej wymienionych kryteriów.
- **Waga kryteriów:** Przydzielenie wagi dla poszczególnych kryteriów, uwzględniając ich znaczenie dla specyfiki działalności wybranej organizacji.
- **Ocena:** Każde narzędzie powinno być ocenione w skali od 1 do 10 dla każdego kryterium. Wynik końcowy dla każdego narzędzia powinien być obliczony jako ważona suma ocen.
- **Rekomendacja:** Na podstawie analizy i macierzy decyzyjnej proszę wybrać i uzasadnić wybór najlepszego narzędzia BI dla analizowanej organizacji.

# Ćwiczenie 13 – E2E Projekt (22.05.2024 – 05.06.2024)

**Zadanie:** End2End Project - Analiza i predykcja kursów walut jako system wspomagania decyzji operacyjnych w firmie zajmującej się tradinigiem walutowym.

**Cel zadania:** Przygotowanie kompleksowego projektu, który będzie analizował i przewidywał kursy walut na podstawie historycznych danych pobieranych z Internetu. Projekt powinien obejmować automatyczny pipeline danych oraz dynamiczny raport, który aktualizuje się w czasie rzeczywistym, w tym:

- **Pobieranie danych:**  
Wybrać i zintegrować API, które dostarcza dane o kursach walut (np. API udostępniane przez bank centralny, Forex lub inne platformy finansowe).
- **Przetwarzanie i przygotowanie danych:**  
Oczyszczenie i przetworzenie surowych danych: usunięcie braków danych, konwersja formatów, normalizacja.  
Przygotowanie danych do analizy i wizualizacji, z uwzględnieniem różnych okien czasowych: lata, rok, miesiąc, tydzień, dzień, godzina.
- **Analiza danych:**  
Analiza trendów kursów walut.  
Wyznaczenie lokalnych minimów i maksimów kursów w danych okresach.
- **Modelowanie predykcyjne:**  
Wykorzystanie modeli uczenia maszynowego do przewidywania przyszłych wartości kursów walut.  
Walidacja modelu za pomocą danych historycznych.
- **Automatyzacja pipeline:**  
Stworzenie automatycznego pipeline'a, który regularnie pobiera, przetwarza dane i aktualizuje model predykcyjny.  
Zastosowanie narzędzi do zarządzania przepływem pracy.
- **Raportowanie i dashboard:**  
Tworzenie interaktywnego dashboardu w narzędziu takim jak Power BI, Tableau lub innych (np. z użyciem bibliotek w Pythonie (np. Dash by Plotly)).  
Dashboard powinien zawierać dynamiczne wykresy i wizualizacje prezentujące aktualne i historyczne kursy walut, a także przewidywania na przyszłość.  
Implementacja KPIs, w tym wskaźnika "kupić/nie kupić", który będzie bazował na przewidywaniach modelu i określonych zasadach (np. kupować, gdy oczekiwany wzrost kursu przekracza 5% - predefiniowane).  
Opcja szczegółowego przeglądu danych dla poszczególnych okien czasowych.

**Wymagania końcowe:** Przygotowanie prezentacji wyników projektu, w której zostaną omówione kluczowe odkrycia, metodyka pracy oraz potencjalne implikacje biznesowe. Projekt powinien być złożony, w pełni funkcjonalny i demonstrujący umiejętności analizy danych, modelowania predykcyjnego oraz budowania raportów i dashboardów. Projekt powinien również zawierać dokumentację opisującą architekturę systemu, użyte narzędzia, procesy oraz kody źródłowe. Detaliczny opis WSZYSTKICH wykonanych kroków i przesłanie je prowadzącemu (wraz z rezultatami, kodami, plikami raportów itd..) do dn. **05.06.2024 (EOD)**. Uwaga – kompleksowa realizacja projektu będzie in(de)krementować ocenę z kolokwium zaliczeniowego (wpływać na Ocenę Końcową).