

Temat 6 .

dr inż. Michał Malinowski

<https://www.drmalinowski.edu.pl>

Rodzaje danych
przechowywanych
w bazach danych

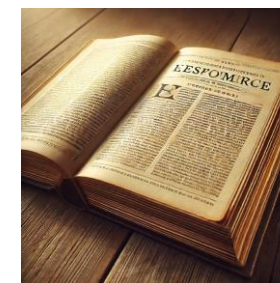
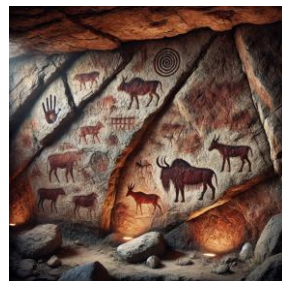
PODSTAWY BAZ DANYCH



Zagadnienia:

- Ilość i nośniki danych na przestrzeni historii
- Formaty wymiany danych
 - CSV
 - XML
 - INI
 - JSON
- Rodzaje danych
 - Dane transakcyjne
 - Logi
 - Dane konfiguracyjne
 - Dane osobowe
 - Dane operacyjne
 - Dane analityczne
 - Dane referencyjne
 - Dane multimedialne
 - Dane geolokalizacyjne
 - Dane tymczasowe

Ilość i nośniki danych na przestrzeni historii



Możliwości gromadzenia i propagowania informacji

Smartfon

2024r

124 GB

> 4 000 000 000
mieszkańców ziemi



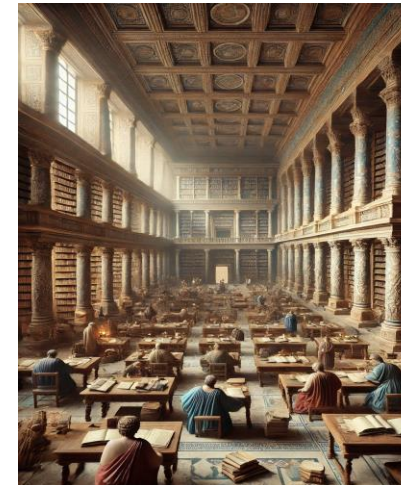
40 000 km
~ 200 ms

Biblioteka Aleksandryjska

II wieku p.n.e.

12,5 GB

Pojemność danych



centrum wiedzy
starożytnego świata

Czas propagacji
informacji



4 000 km
~ 11 dnia

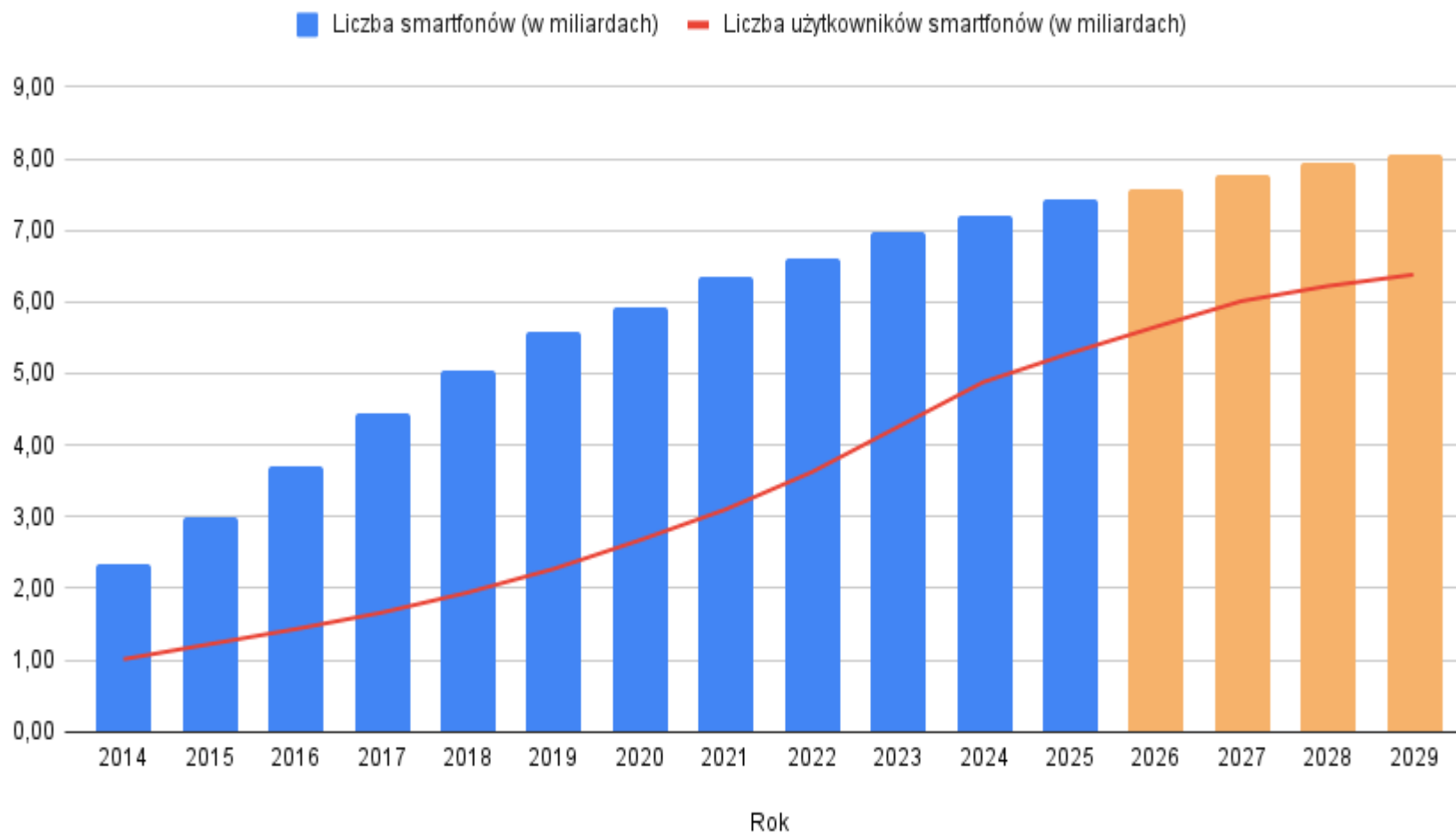
Media propagacji informacji

Dystans 4 000 km



Statystyka smartfonów 2014-2029

Statystyka smartfonów 2014 - 2029



Formaty wymiany danych

Formaty wymiany danych

to standardowe struktury i sposoby zapisu informacji, które umożliwiają interoperacyjność między różnymi systemami bazodanowymi, aplikacjami oraz narzędziami.

W kontekście baz danych pełnią kluczową rolę w **przechowywaniu, eksportowaniu, importowaniu** oraz **integracji danych** w sposób, który zachowuje ich spójność, strukturę i znaczenie

Cechy formatów wymiany danych

Uniwersalność

- Powinien być obsługiwany przez wiele technologii i platform (np. systemy relacyjne i nierelacyjne).

Elastyczność

- Umożliwiają reprezentację danych w postaci tabel, hierarchii lub złożonych struktur (np. obiektów).

Czytelność

- Format powinien być czytelny zarówno dla ludzi, jak i dla systemów komputerowych.

CSV

CSV (Comma-Separated Values)

- Struktura: Format tekstowy reprezentujący dane w formie tabelarycznej. Poszczególne wartości są oddzielone przecinkami (lub innymi separatorami), a każdy wiersz odpowiada jednemu rekordowi.
- Czytelność: Czytelny dla ludzi i prosty do przetwarzania, ale nie obsługuje hierarchii czy złożonych struktur danych.
- Użycie: Często stosowany do eksportu i importu danych z baz danych, arkuszy kalkulacyjnych i innych narzędzi analitycznych.

CSV

Budowa

```
Imię i nazwisko,Adres email,Numer telefonu,Miejsce zamieszkania  
"Jan Kowalski",jankowalski@example.com,555-1234,Warszawa  
"Anna Nowak",annanowak@example.com,555-5678,Kraków  
"Paweł Wiśniewski",pawelwisniewski@example.com,555-9012,Wrocław
```

- Plik tekstowy z rozszerzeniem .CSV
- Pierwsza linia to nagłówek, który definiuje nazwy kolumn.
- Każda kolejna linia reprezentuje jeden rekord, czyli jedną osobę z jej danymi kontaktowymi.
- Wartości są oddzielone przecinkami.
- Każda linia kończy się znakiem nowej linii, oddzielając rekordy od siebie.

XML

XML (eXtensible Markup Language)

- Struktura: Format oparty na znacznikach, pozwalający na definiowanie niestandardowych tagów, co zapewnia dużą elastyczność. Struktura dokumentu XML jest reprezentowana przez drzewo.
- Czytelność: Czytelny dla ludzi, ale mniej zwarty niż JSON.
- Użycie: Używany w wielu systemach do przechowywania i transportu danych, w usługach webowych, w dokumentach (np. Microsoft Office) oraz w konfiguracjach.

XML

Budowa

```
<?xml version="1.0" encoding="UTF-8"?>
<kontakty>
  <osoba>
    <imie_i_nazwisko>Jan Kowalski</imie_i_nazwisko>
    <adres_email>jankowalski@example.com</adres_email>
    <numer_telefonu>555-1234</numer_telefonu>
    <miejsce_zamieszkania>Warszawa</miejsce_zamieszkania>
  </osoba>
  <osoba>
    <imie_i_nazwisko>Anna Nowak</imie_i_nazwisko>
    <adres_email>annanowak@example.com</adres_email>
    <numer_telefonu>555-5678</numer_telefonu>
    <miejsce_zamieszkania>Kraków</miejsce_zamieszkania>
  </osoba>
</kontakty>
```

- Plik tekstowy z rozszerzeniem .xml
- Pierwsza linijka to deklaracja informuje parser, że dokument jest w formacie XML wersji 1.0 i używa kodowania UTF-8.
- Każdy kontakt jest reprezentowany przez element <osoba>.
- Dane są wydzielone do odpowiednich elementów potomnych, takich jak <imie_i_nazwisko>, <adres_email>, <numer_telefonu>, i <miejsce_zamieszkania>.
- Elementy te są zagnieżdżone wewnątrz głównego elementu <kontakty>, który zawiera całą listę kontaktów.

INI

INI (Initialization File Format)

- Struktura: Format tekstowy używany do przechowywania konfiguracji w prostych parach klucz-wartość, z danymi zgrupowanymi w sekcje. Sekcje są zaznaczone w nawiasach kwadratowych [], a każda para klucz-wartość jest oddzielona znakiem równości =.
- Czytelność: Format INI jest łatwy do odczytania i edycji przez ludzi, z wyraźnie oddzielonymi sekcjami i prostymi parametrami konfiguracji. Nie obsługuje jednak zagnieżdżonych struktur danych ani bardziej złożonych typów danych.
- Użycie: Format INI jest szeroko stosowany w konfiguracji oprogramowania, szczególnie na platformach Windows, ale także w różnych aplikacjach na innych systemach operacyjnych. Jest też często używany w starszym oprogramowaniu i grach komputerowych.

INI

Budowa

```
; Komentarz - zaczyna się od znaku średnika

[OpcjeAplikacji]
klucz1 = wartość1
klucz2 = wartość2
klucz3 = wartość3

[InnaSekcja]
ustawienieA = true
ustawienieB = 42
```

- Plik tekstowy z rozszerzeniem .ini
- Linie zaczynające się od znaku średnika (;) są komentarzami i są ignorowane podczas analizy.
- Sekcje są otoczone nawiasami kwadratowymi ([nazwa_sekcji]) i zawierają pary klucz-wartość związanymi z tą sekcją.
- Każda para klucz-wartość składa się z klucza (po lewej stronie znaku równości =) i wartości (po prawej stronie znaku równości =).

JSON

JSON (JavaScript Object Notation)

- Struktura: Lekki format tekstowy, bazujący na składni obiektów JavaScript. Używa nawiasów klamrowych {} do reprezentacji obiektów (czyli zestawów par klucz-wartość) i nawiasów kwadratowych [] dla tablic.
- Czytelność: Bardzo czytelny zarówno dla ludzi, jak i maszyn.
- Użycie: Szeroko stosowany w aplikacjach internetowych dla wymiany danych między klientem a serwerem, a także jako format konfiguracyjny i w web APIs.

JSON

Budowa

```
{
  "imie": "Jan",
  "nazwisko": "Kowalski",
  "wiek": 30,
  "miasto": "Warszawa",
  "znajomi": ["Anna", "Paweł", "Ewa"],
  "adres": {
    "ulica": "Krakowska 123",
    "kod_pocztowy": "00-000",
    "kraj": "Polska"
  }
}
```

- Plik tekstowy z rozszerzeniem .json
- Dane są przedstawione w formie par "klucz"-"wartość".
- Obiekty (zestawy danych) są otoczone klamrami {}.
- Klucze są zawsze w podwójnych cudzysłowach (np. "imie").
- Wartości mogą być typu tekstowego, liczbowego, logicznego, tablicy lub innego obiektu JSON.
- Tablice są listami wartości otoczone kwadratowymi nawiasami [].

Formaty wymiany danych

Porównanie

Format	Najlepsze zastosowanie	Główne zalety	Ograniczenia
CSV	Tabelaryczne dane (import/eksport).	Prosty, lekki, łatwy do odczytu.	Brak wsparcia dla zagnieżdżonych struktur.
JSON	Złożone struktury danych (np. dokumenty, API).	Lekki, wsparcie dla zagnieżdżenia.	Mniej efektywny dla danych tabelarycznych.
XML	Złożone dane wymagające walidacji (np. dokumenty).	Wszechstronny, walidacja danych.	Duża objętość, trudniejszy w obsłudze.
INI	Pliki konfiguracyjne.	Prosty, czytelny.	Ograniczone możliwości przechowywania danych.

Rodzaje danych

Rodzaj danych w kontekście baz danych

to klasyfikacja informacji przechowywanych w systemach bazodanowych, oparta na charakterze, strukturze, przeznaczeniu oraz zastosowaniach tych danych w procesach biznesowych i technologicznych.

Każdy rodzaj danych reprezentuje specyficzną grupę informacji, które mogą być różnie przetwarzane, analizowane i przechowywane w zależności od ich właściwości i wymagań.

Pomagają w organizacji i optymalizacji systemów bazodanowych, zapewniając efektywność w przechowywaniu, analizie i wymianie informacji między systemami.

Cechy definiujące rodzaje danych

Przeznaczenie

- Dane są klasyfikowane według ich roli w systemie, np. dane transakcyjne obsługują operacje finansowe, a dane osobowe dotyczą użytkowników.

Struktura

- Określa, czy dane mają strukturę tabelaryczną, hierarchiczną, dokumentową czy prostą lub multimedialną.

Format i sposób przechowywania

- Rodzaje danych determinują format przechowywania (np. relacyjne, dokumentowe) i typy baz danych używanych do ich zarządzania.

Charakterystyka wzrostu

- Różne rodzaje danych charakteryzują się różną dynamiką przyrostu (np. dane logów mogą mieć gwałtowny wzrost, podczas gdy dane konfiguracyjne rozwijają się powoli).

Przykładowe zastosowania

- Określają, w jakich procesach biznesowych lub technologicznych dany rodzaj informacji znajduje głównie zastosowanie.

Rodzaje danych

Rodzaj danych	Opis	Przykłady
Dane transakcyjne	Informacje finansowe i biznesowe związane z procesami transakcji, wymagające spójności i integralności (ACID).	Operacje bankowe, płatności kartowe, zamówienia w systemach e-commerce.
Logi	Rejestry zdarzeń systemów, aplikacji, urządzeń; mogą mieć różną strukturę i duży wolumen.	Pliki logów serwerów www, logi systemowe (syslog), logi aplikacji (zdarzenia, błędy).
Dane konfiguracyjne	Zbiór par klucz–wartość lub definicji wspomagających działanie aplikacji oraz ułatwiających standaryzację danych.	Konfiguracje systemowe (klucz → wartość), mapy klucz–definicja, parametry uruchomieniowe aplikacji.
Dane osobowe (dane użytkowników)	Informacje o osobach (klientach, pracownikach, użytkownikach systemów), często wymagające ochrony (RODO/GDPR) i zarządzania tożsamościami (IAM).	Imię i nazwisko, adres, numery kontaktowe, hasze haseł, tokeny uwierzytelniające, uprawnienia użytkownika.
Dane operacyjne (produkcyjne)	Dane przetwarzane na bieżąco w procesach biznesowych.	Rekordy produktów, przepływy pieniężne, aktywne zamówienia, śledzenie przesyłek, wypożyczenia filmów.
Dane analityczne (OLAP)	Zbiory danych przeznaczone do raportowania, analiz wielowymiarowych i BI (Business Intelligence).	Zestawienia sprzedaży, dane historyczne, raporty menedżerskie, zagregowane dane o klientach.
Dane referencyjne	Słowniki i zbiory pomocnicze do standaryzacji i ujednolicania innych danych.	Listy krajów, województw i miast, kody pocztowe, kody jednostek miar, kategorie produktów.
Dane multimedialne	Bardzo duże pliki binarne lub strumienie multimedialne – obrazy, audio, wideo.	Zdjęcia w mediach społecznościowych, biblioteki muzyczne, archiwa wideo, streaming wideo na żywo.
Dane geolokalizacyjne (GIS)	Dane przestrzenne, opisujące lokalizację obiektów w przestrzeni 2D lub 3D i ich atrybuty związane z lokalizacją	Koordynaty GPS, mapy, warstwy GIS.
Dane czasoszeręgowe	Rekordy oznaczone czasem, nastawione na analizę trendów i agregację w czasie.	Dane z czujników IoT, metryki systemowe (CPU, pamięć), notowania giełdowe, logi (gdy kluczowy jest aspekt czasowy).
Dane streamingowe	Dane napływające w trybie ciągłym z wielu źródeł (zdarzenia, logi, sensory).	Ruch sieciowy w systemach telemetrycznych, zdarzenia z aplikacji, informacje z sensorów przemysłowych.
Dane nieustrukturyzowane	Informacje bez ściśle zdefiniowanej struktury tabel (np. tekst, e-maile, załączniki).	E-maile, dokumenty tekstowe, komentarze na forach, dane w różnych formatach (PDF, DOCX).
Dane semiestrukturyzowane	Dane częściowo zdefiniowane, zazwyczaj w formatach XML/JSON, elastyczne pod względem schematu.	Wiadomości XML, pliki JSON, eksporty danych z systemów ERP i aplikacji.
Dane konfiguracyjne	Informacje niezbędne do poprawnego działania systemów i aplikacji.	Pliki konfiguracyjne (YAML, JSON), parametry aplikacji, definicje usług w architekturze mikroserwisowej.
Dane tymczasowe	Dane o krótkiej żywotności, przechowywane tymczasowo.	Sesje użytkowników, tokeny autoryzacyjne, kolejki tymczasowych zadań, koszyki zakupowe lub cookie.
Metadane	Dane opisujące i charakteryzujące inne dane (tzw. „dane o danych”).	Opisy plików (format, autor, data), definicje tabel i pól w bazach relacyjnych, słowniki danych w hurtowniach.
Dane treningowe AI (korpusy uczące)	Zbiory danych wykorzystywane do uczenia maszynowego oraz wektory cech (embeddings) używane w systemach AI (wyszukiwanie semantyczne, rekomendacje).	Korpusy tekstów dla NLP, obrazy z etykietami dla sieci neuronowych, wektorowe reprezentacje dokumentów/obrazów (umieszczane np. w bazach wektorowych).
Dane z systemów IoT	Dane generowane przez urządzenia i sensory w ramach Internetu Rzeczy, obejmujące pomiary, stany maszyn i sygnały sterujące.	Temperatury, ciśnienia, wibracje w liniach produkcyjnych, komunikaty diagnostyczne, pozycje obiektów ruchomych (floty pojazdów, dronów).

Dane transakcyjne

Dane transakcyjne

Informacje finansowe i biznesowe związane z procesami transakcji, wymagające spójności i integralności (ACID).

Przykład

Operacje bankowe, płatności kartowe, zamówienia w systemach e-commerce.

ID_transakcji	Data_transakcji	Typ_transakcji	Kwota	Waluta	Status_transakcji
101	20.12.2024	Płatność kartą	250.5	PLN	Zakończona
102	21.12.2024	Przelew bankowy	1200.0	PLN	Zakończona
103	22.12.2024	Zakup online	75.99	PLN	Oczekująca
104	23.12.2024	Wpłata gotówkowa	500.0	PLN	Zakończona
105	23.12.2024	Płatność			

Typowe bazy / narzędzia

Bazy relacyjne

NoSQL (np. kolumnowe, klucz–wartość) w aplikacjach o bardzo dużej skali.

ID_transakcji	Klient	Produkt	Kwota	Data_zakupu	Metoda_płatności	Status
TRX001	Jan Kowalski	Laptop	3500.0	01.12.2024	Karta kredytowa	Zrealizowana
TRX002	Anna Nowak	Smartfon	2800.0	02.12.2024	Blik	Zrealizowana
TRX003	Piotr Wiśniewski	Książka	50.0	03.12.2024	PayPal	Oczekująca
TRX004	Maria Kowalczyk	Abonament streamingowy	25.99	03.12.2024	Przelew bankowy	Zrealizowana
TRX005	Tomasz Lewandowski	Akcesoria do telefonu	120.0	04.12.2024	Płatność mobilna	Anulowana

Dane transakcyjne

Przyrost danych

Najczęściej stały, liniowy wzrost (**5–15% rocznie**), zależny od liczby klientów i transakcji.

$$P = N \times D \times A$$

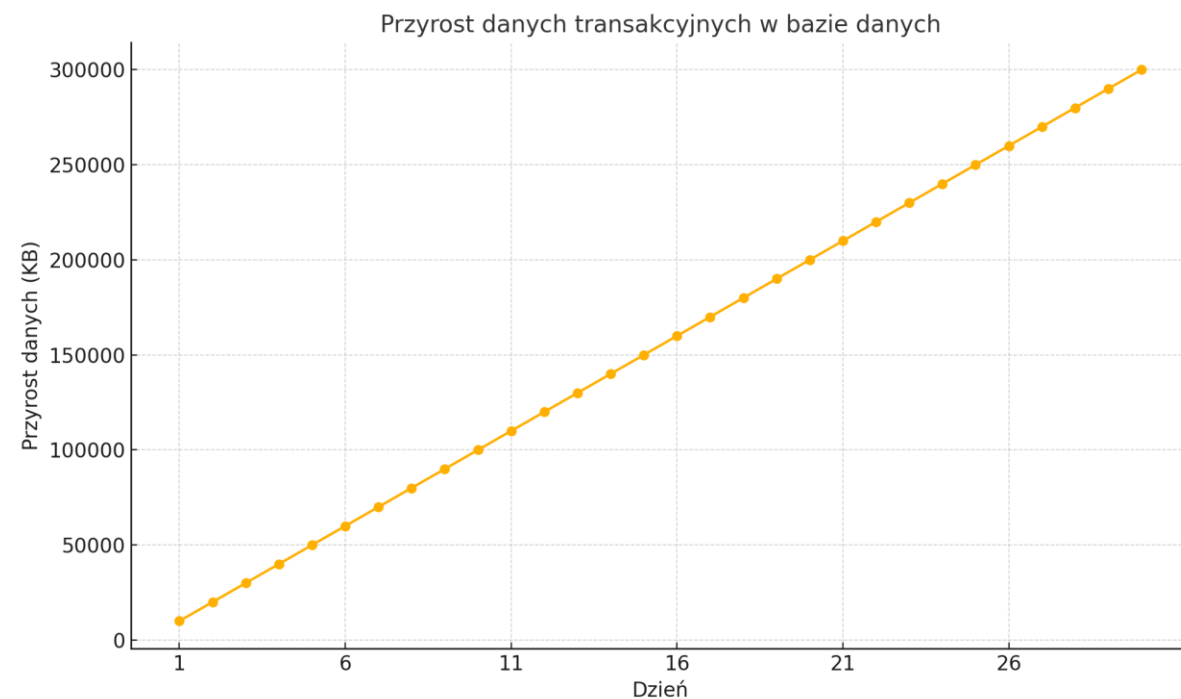
Gdzie:

P – całkowity przyrost danych w określonym przedziale czasu,

N – średnia liczba nowych transakcji w okresie,

D – liczba okresów (np. dni, miesięcy),

A – średni rozmiar jednej transakcji (np. w KB).



Szacowany udział

25–30%

Logi

Logi

Rejestry zdarzeń systemów, aplikacji, urządzeń; mogą mieć różną strukturę i duży wolumen.

Przykład

Pliki logów serwerów www, logi systemowe (syslog), logi aplikacji (zdarzenia, błędy).

Typowe bazy / narzędzia

Bazy dokumentowe (JSON, XML)

Bazy kolumnowe

Bazy klucz–wartość

Narzędzia do analizy logów (Elasticsearch, Splunk)

```
[
  {
    "Czas": "2024-12-20 15:30:45",
    "Typ_logu": "Serwer HTTP",
    "Adres_IP": "192.168.1.10",
    "Działanie": "Otrzymano żądanie POST /login",
    "Status": "Sukces"
  },
  {
    "Czas": "2024-12-20 15:31:10",
    "Typ_logu": "Firewall",
    "Adres_IP": "10.0.0.5",
    "Działanie": "Zablokowano próbę nieautoryzowanego dostępu",
    "Status": "Zablokowano"
  },
  {
    "Czas": "2024-12-20 15:32:05",
    "Typ_logu": "Baza danych",
    "Adres_IP": "172.16.0.3",
    "Działanie": "Zapisano nowy rekord w tabeli 'uzytkownicy'.",
    "Status": "Sukces"
  }
]
```

```
<logi>
  <log>
    <czas>2024-12-20T16:00:01</czas>
    <typ_logu>Antywirus</typ_logu>
    <nazwa_stacji>Stacja001</nazwa_stacji>
    <zdarzenie>Wykryto zagrożenie: Trojan.Win32.Generic</zdarzenie>
    <status>Zneutralizowano</status>
  </log>
  <log>
    <czas>2024-12-20T16:05:45</czas>
    <typ_logu>Antywirus</typ_logu>
    <nazwa_stacji>Stacja002</nazwa_stacji>
    <zdarzenie>Przeprowadzono pełne skanowanie systemu</zdarzenie>
    <status>Zakończono pomyślnie</status>
  </log>
  <log>
    <czas>2024-12-20T16:10:30</czas>
    <typ_logu>Antywirus</typ_logu>
    <nazwa_stacji>Stacja003</nazwa_stacji>
    <zdarzenie>Wykryto próbę dostępu do zainfekowanego pliku</zdarzenie>
    <status>Zablokowano</status>
  </log>
</logi>
```

Logi

Przyrost danych

Często wysoki, wprost proporcjonalny do ruchu;
w dynamicznych środowiskach może rosnać nawet **>50% rocznie**.

$$P=R \times D \times S$$

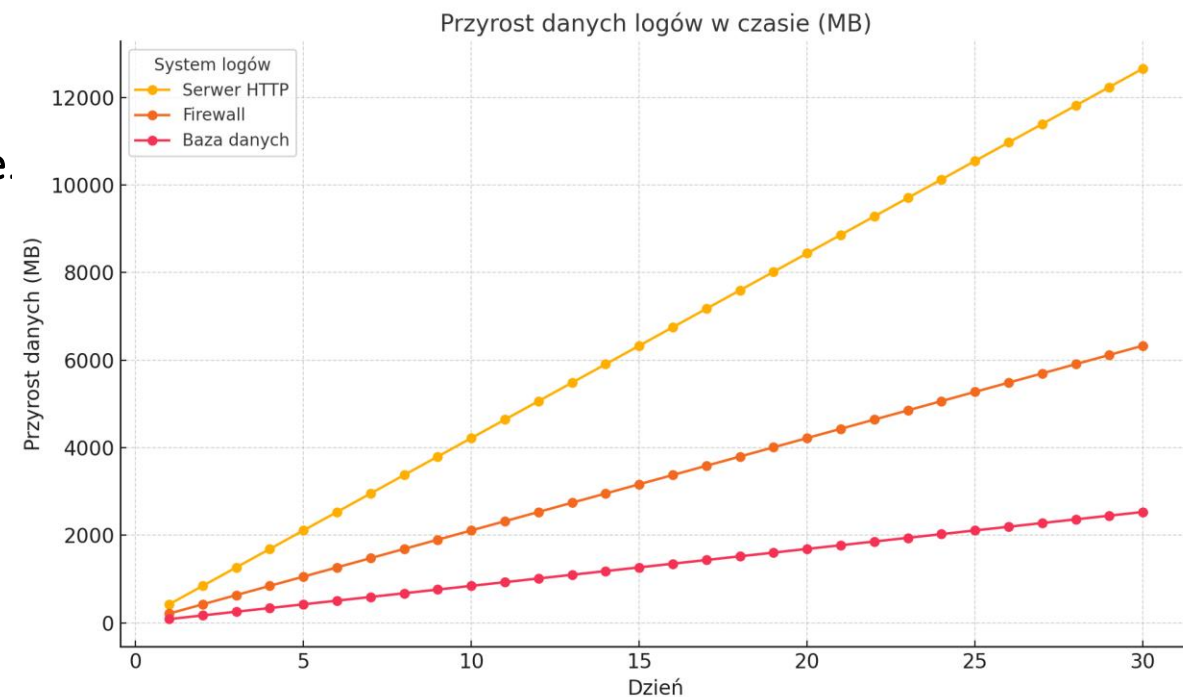
Gdzie:

P – całkowity przyrost danych w określonym przedziale czasu,

R – średnia liczba zdarzeń logów w okresie (np. żądań HTTP, wpisów w logach systemowych),

D – liczba okresów (np. dni, miesiące),

S – średni rozmiar pojedynczego wpisu w logu (np. w KB).



Szacowany udział

40–60%

Dane konfiguracyjne

Dane konfiguracyjne

Zbiór par klucz–wartość lub definicji wspomagających działanie aplikacji oraz ułatwiających standaryzację danych.

Przykład

Konfiguracje systemowe (klucz → wartość), parametry uruchomieniowe aplikacji.

Typowe bazy / narzędzia

Bazy klucz–wartość bazy dokumentowe,
Bazy dokumentowe gdy konfiguracja ma bardziej rozbudowaną strukturę.

```
[Database]
Type=Document
Engine=MongoDB
Host=127.0.0.1
Port=27017
Authentication=Yes

[Authentication]
Username=admin
Password=SuperSecurePassword
Mechanism=SCRAM-SHA-256

[ReplicaSet]
Enabled=True
ReplicaSetName=rs0
Members=3
Arbiter=No

[Storage]
Engine=WiredTiger
Compression=zlib
Journal=True

[Network]
SSID=MyHomeWiFi
Mode=802.11ac
Channel=6
Bandwidth=20MHz

[Security]
Encryption=WPA2-PSK
Password=SuperSecurePassword123

[LAN]
IP=192.168.0.1
SubnetMask=255.255.255.0
DHCP=Enabled

[WAN]
ConnectionType=PPPoE
Username=myISPuser
Password=myISPpassword

[Firewall]
Enabled=True
AllowedPorts=80, 443, 22
DefaultPolicy=Drop
```

Dane konfiguracyjne

Przyrost danych

Niewielki, skokowy wzrost (**2-5 % rocznie**). Rośnie głównie wraz z rozbudową systemów i pojawieniem się nowych konfiguracji.

$$P=U \times S \times A$$

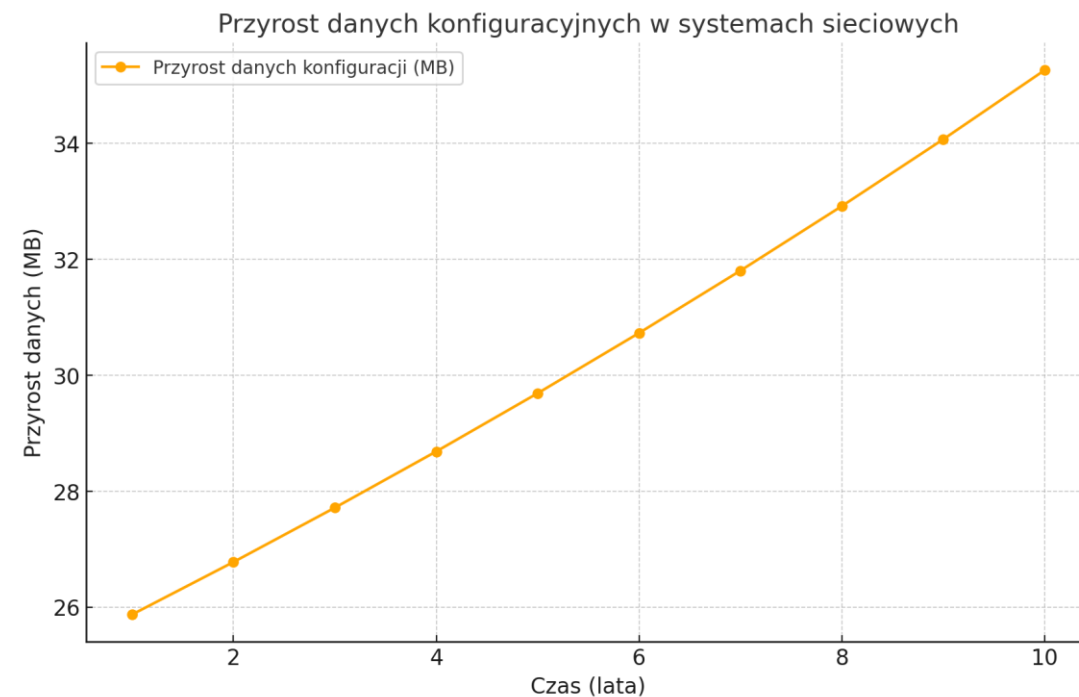
Gdzie:

P – całkowity przyrost danych w określonym przedziale czasu,

U – średnia liczba urządzeń (np. routerów, punktów dostępowych, serwerów),

D – liczba okresów (np. dni, miesiące),

S – średni rozmiar konfiguracji dla jednego urządzenia (np. w KB).



Szacowany udział

5–10%

Dane osobowe

Dane osobowe (użytkowników)

Informacje o osobach (klientach, pracownikach, użytkownikach), wymagające ochrony (RODO/GDPR) i zarządzania tożsamościami.

Przykład

Imię i nazwisko, adres, numery kontaktowe, hasze haseł, PESEL, uprawnienia użytkownika.

Typowe bazy / narzędzia

Bazy Relacyjne (np. bazy CRM/ERP, HR)

LDAP / Active Directory

Usługi IAM (Identity and Access Management)

```
dn: cn=Jan Kowalski,ou=Pracownicy,dc=przyklad,dc=com
objectClass: inetOrgPerson
objectClass: organizationalPerson
objectClass: person
objectClass: top
cn: Jan Kowalski
sn: Kowalski
givenName: Jan
mail: jan.kowalski@przyklad.com
telephoneNumber: +48 123 456 789
title: Specjalista IT
department: Dział Technologii
manager: cn=Anna Nowak,ou=Pracownicy,dc=przyklad,dc=com
userPassword: {SHA}X03M01qnZdYdgyfeuILPmQ==
uid: jkowalski
employeeNumber: 12345
```

Dane osobowe

Przyrost danych

Stopniowy wzrost w miarę rozwoju bazy użytkowników (**10–20% rocznie**), zależny od pozyskiwania nowych klientów lub pracowników.

$$P=U \times D \times S$$

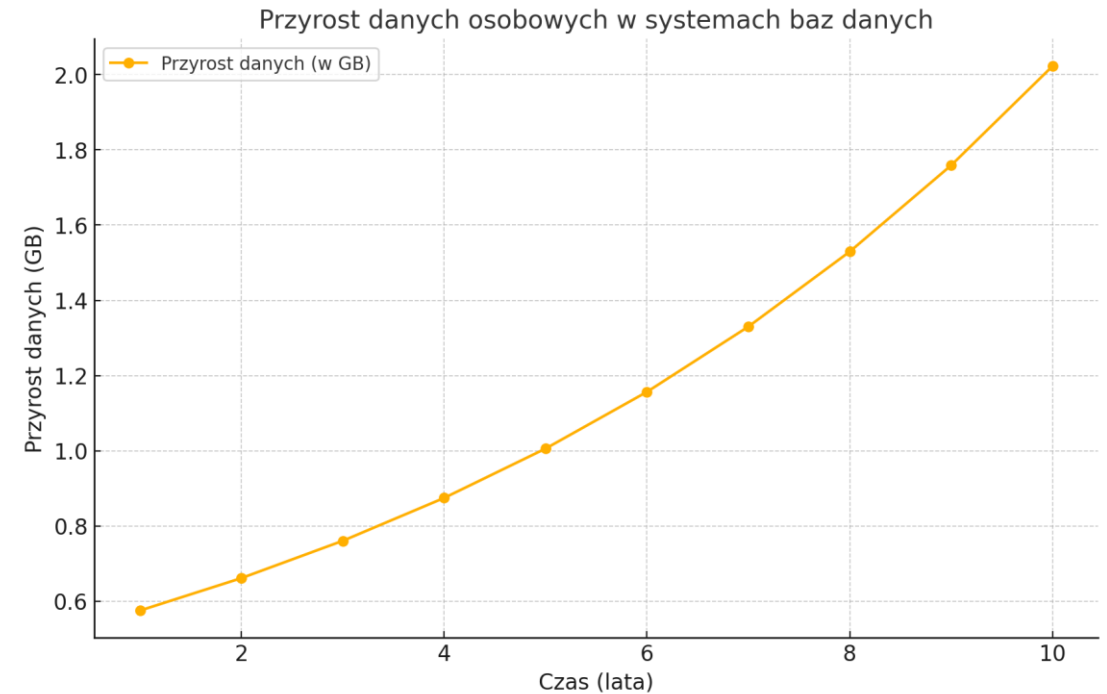
Gdzie:

P – całkowity przyrost danych w określonym przedziale czasu,

U – średnia liczba nowych użytkowników w okresie,

D – liczba okresów (np. dni, miesiące),

S – średni rozmiar danych przypadający na jednego użytkownika.



Szacowany udział

20–25%

Dane operacyjne

Dane operacyjne (produkcyjne)

Dane przetwarzane na bieżąco w procesach biznesowych.

Przykład

Rekordy produktów, przepływy pieniężne, aktywne zamówienia, śledzenie przesyłek, wypożyczenia filmów.

Typowe bazy / narzędzia

Relacyjne bazy transakcyjne
NoSQL (dokumentowe, klucz–wartość) w aplikacjach o bardzo dużej skali

ID Produktu	Nazwa Produktu	Kategoria	Cena (PLN)	Stan Magazynowy	Dostawca	Opis	Data Dodania
PROD001	Laptop Dell XPS 13	Elektronika	5500.00	15	Dell Polska	Lekki laptop z ekranem 13 cali.	10.01.2024
PROD002	Telefon Samsung Galaxy S23	Elektronika	3600.00	25	Samsung Electronics	Smartfon z aparatem 50 MP i 5G.	15.03.2024
PROD003	Monitor LG 27UL850	Elektronika	1800.00	8	LG Electronics	Monitor 4K UHD z HDR i USB-C.	20.02.2024
PROD004	Krzesło biurowe ErgoChair	Meble	750.00	50	ErgoMeble	Ergonomiczne krzesło z regulacją.	05.04.2024
PROD005	Książka "Python w praktyce"	Książki	120.00	100	Helion	Praktyczny przewodnik po Pythonie.	01.05.2024
PROD006	Ekspres do kawy Philips	AGD	1200.00	30	Philips	Automatyczny ekspres do kawy z młynkiem.	25.01.2024
PROD007	Plecak Nike Elemental	Akcesoria	180.00	70	Nike Polska	Pojemny plecak sportowy 20L.	12.06.2024

Dane operacyjne

Przyrost danych

Płynny wzrost (**5–20% rocznie**), zależny od skali i dynamiki procesów biznesowych.

$$P=U \times D \times S$$

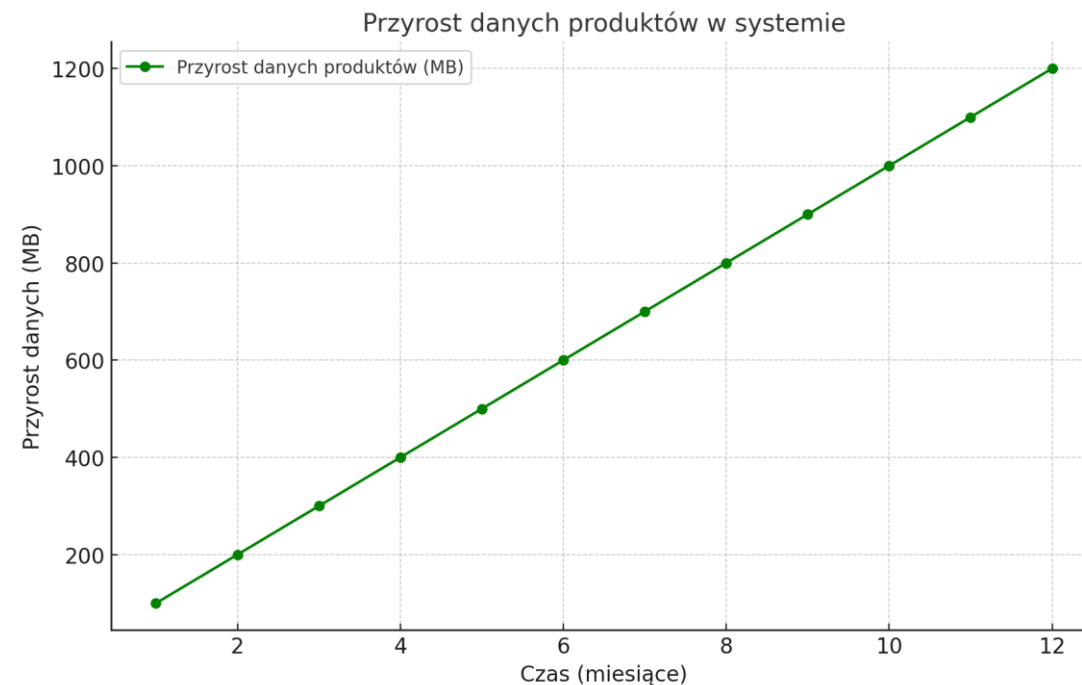
Gdzie:

P – całkowity przyrost danych w określonym przedziale czasu,

U – średnia liczba nowych produktów (zamówień, wypożyczeń) w okresie,

D – liczba okresów (np. dni, miesiące),

S – średni rozmiar danych produktów (zamówienia, wypożyczenia) (np. w KB).



Szacowany udział

30–40%

Dane analityczne

Dane analityczne (OLAP)

Zbiory danych przeznaczone do raportowania, analiz wielowymiarowych i BI (Business Intelligence).

Przykład

Zestawienia sprzedaży, dane historyczne, raporty menedżerskie, zagregowane dane o klientach.

Typowe bazy / narzędzia

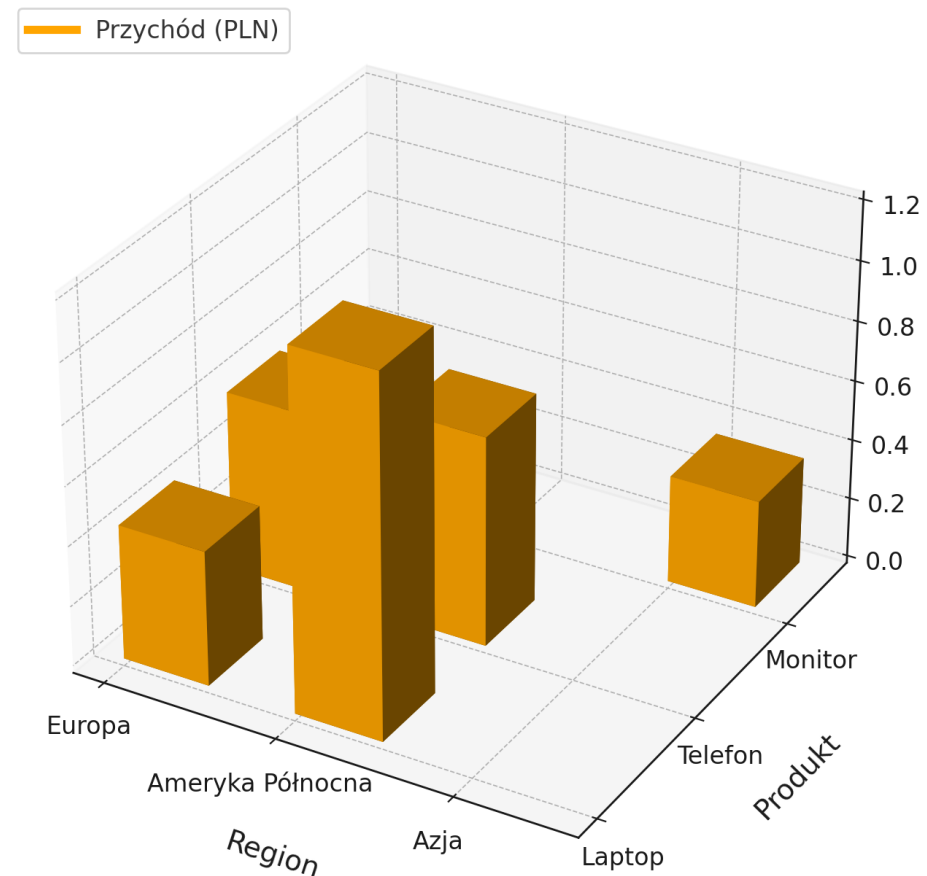
Hurtownie danych

Bazy kolumnowe

Bazy relacyjne z OLAP (ang. Online Analytical Processing)

OLAP to technologia przetwarzania danych, która umożliwia szybkie analizowanie dużych zbiorów danych w wielu wymiarach

Przychód w podziale na regiony i produkty (3D)



Dane analityczne

Przyrost danych

Często kumulatywny i długookresowy wzrost; rozbudowa historycznych danych może sięgać **20–30% rocznie**.

$$P = (TxST + NxSN) \times D$$

Gdzie:

P – całkowity przyrost danych w określonym przedziale czasu,

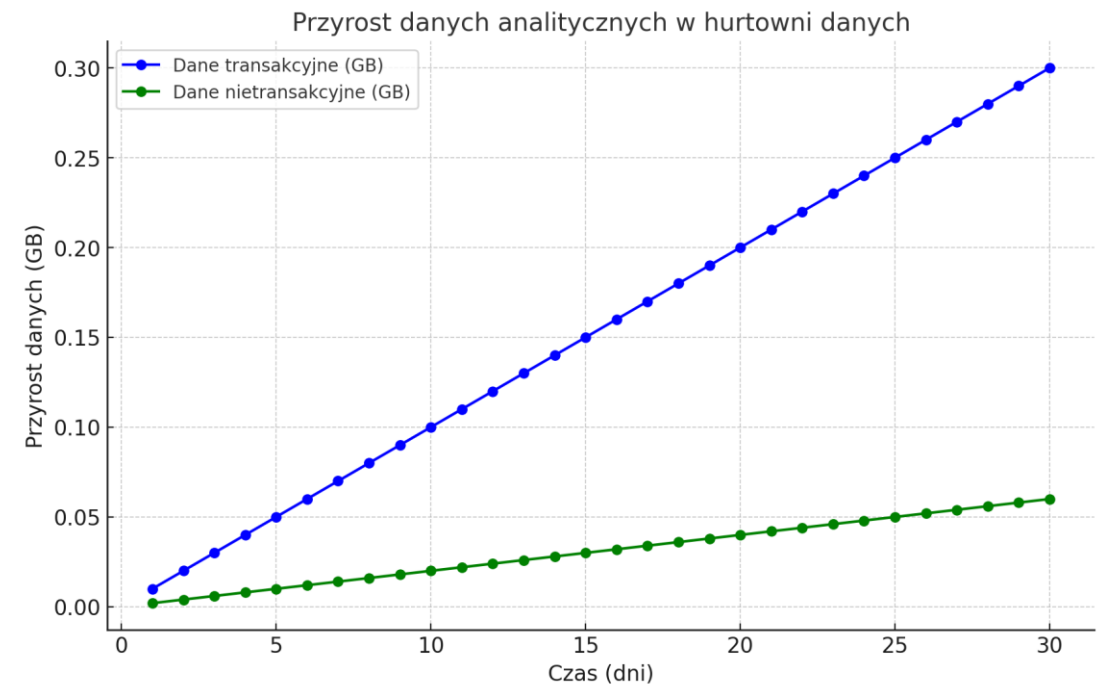
T – liczba nowych transakcji lub rekordów dodawanych do hurtowni w okresie,

N – liczba nowych danych nietransakcyjnych w okresie,

D – liczba okresów (np. dni, miesiące),

ST – średni rozmiar danych transakcyjnych

SN – średni rozmiar danych nietransakcyjnych .



Szacowany udział

15–20%

Dane referencyjne

Dane referencyjne

Słowniki i zbiory pomocnicze do standaryzacji i ujednolicania innych danych.

Przykład

Listy krajów, województw i miast, kody pocztowe, kody jednostek miar, kategorie produktów.

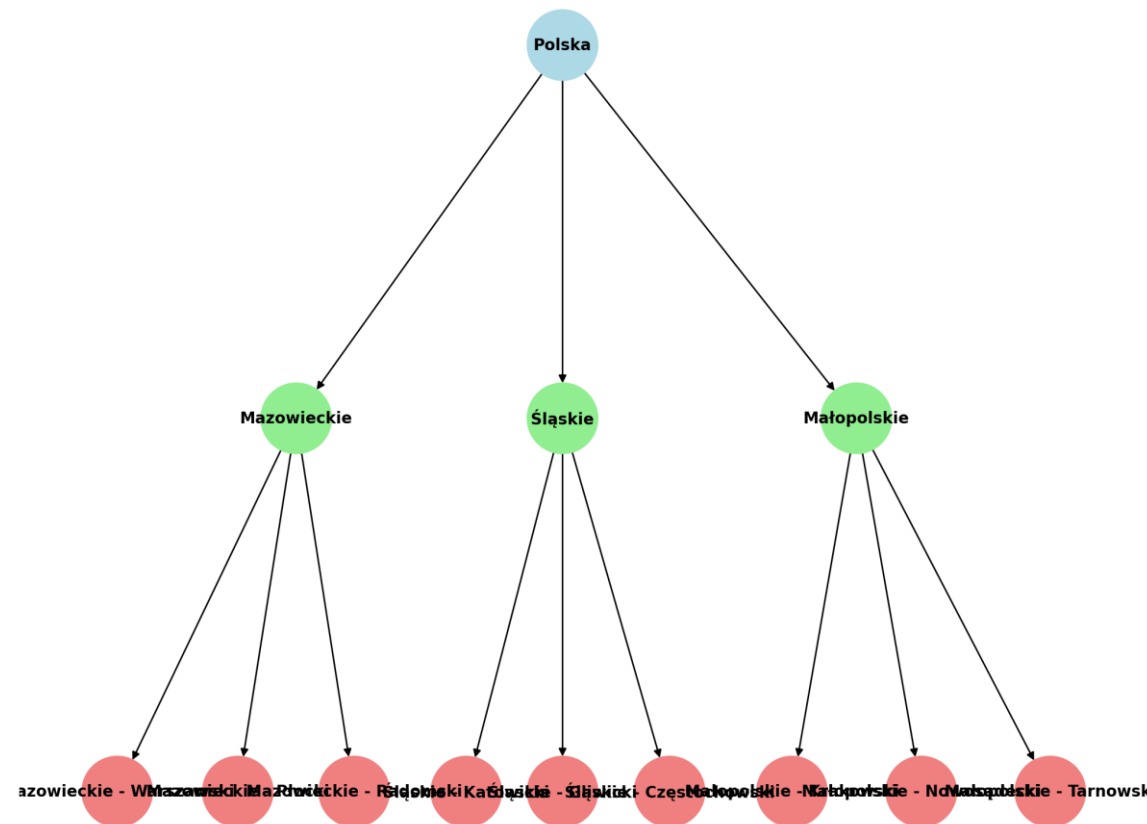
Typowe bazy / narzędzia

Bazy relacyjne

Bazy grafowe aby oddać hierarchię

Bazy klucz–wartość, gdy potrzebna jest szybka dostępność w środowiskach rozproszonych.

Hierarchia administracyjna: Polska > Wybrane Województwa > Powiaty



Dane referencyjne

Przyrost danych

Niewielki przyrost (zwykle <5% rocznie); duże zmiany częściej wynikają z restrukturyzacji niż z naturalnego wzrostu.

$$P = U \times (R1 - R2) \times S \times D$$

Gdzie:

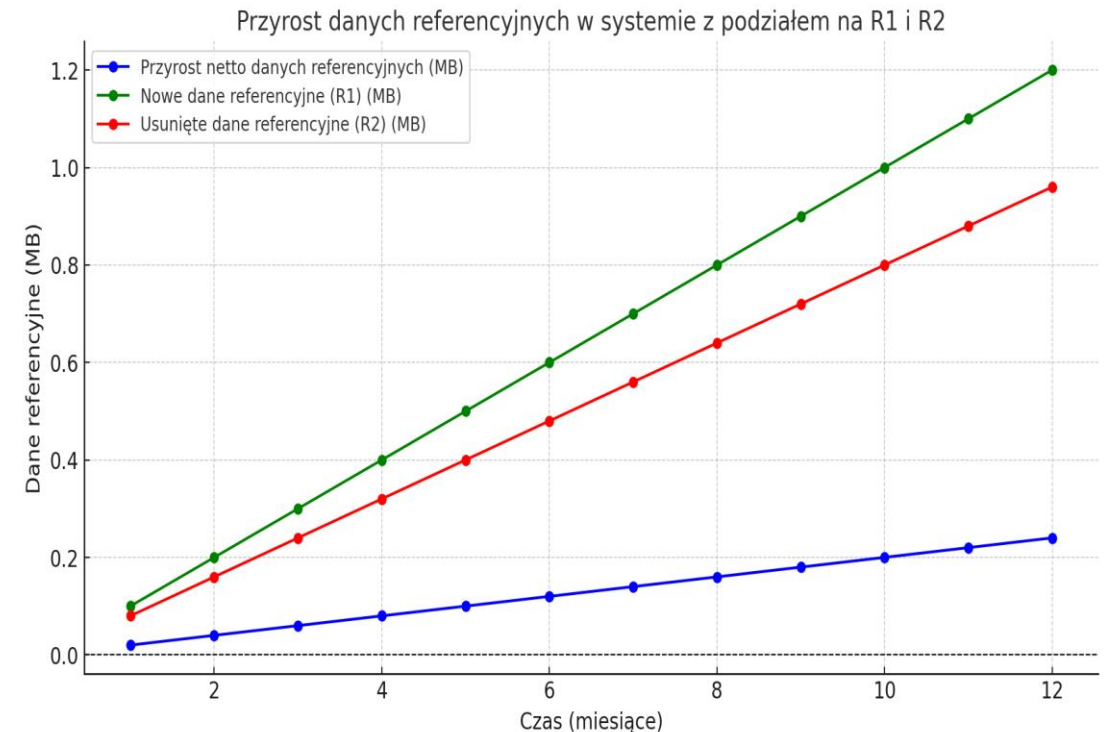
P – całkowity przyrost danych w określonym przedziale czasu,

R1 – średnia liczba nowych danych referencyjnych w okresie,

R2 – średnia liczba usuniętych danych referencyjnych w okresie,

D – liczba okresów (np. dni, miesiące),

S – średni rozmiar danych referencyjnych



Szacowany udział

5–10%

Dane multimedialne

Dane multimedialne

Bardzo duże pliki binarne lub strumienie multimedialne –
obrazy, audio, wideo.

Przykład

Zdjęcia w mediach społecznościowych, biblioteki
muzyczne, archiwa wideo, streaming wideo na żywo.

Typowe bazy / narzędzia

Obiektowe magazyny plików

Metadane w bazach relacyjnych (zawartość w systemie plików)

Bazy NoSQL

ID Filmu	Tytuł	Rok	Gatunek	Reżyser	Oce na	Czas Trwania (min)	Wielkość Pliku (GB)	Plik Filmu
1	Skazani na Shawshank	1994	Dramat	Frank Darabont	9,3	142	1,40	shawshank_redemption.mp4
2	Ojciec Chrzestny	1972	Kryminat	Francis Ford Coppola	8,8	175	1,50	godfather.mp4
3	Mroczny Rycerz	2008	Akcja	Christopher Nolan	7,5	152	2,00	dark_knight.mp4
4	Pulp Fiction	1994	Kryminat	Quentin Tarantino	8,9	154	1,20	pulp_fiction.mp4
5	Władca Pierścieni: Powrót Króla	2003	Fantasy	Peter Jackson	9,3	201	0,55	return_of_the_king.mp4

System plików

shawshank_redemption.mp4

1,4 GB

godfather.mp4

1,5 GB

1,5 GB

dark_knight.mp4

Dane multimedialne

Przyrost danych

Bardzo duży przyrost objętościowy (rozmiary plików), w popularnych serwisach może rosnać wykładniczo (**około 30% rocznie**).

$$P=N \times S \times D$$

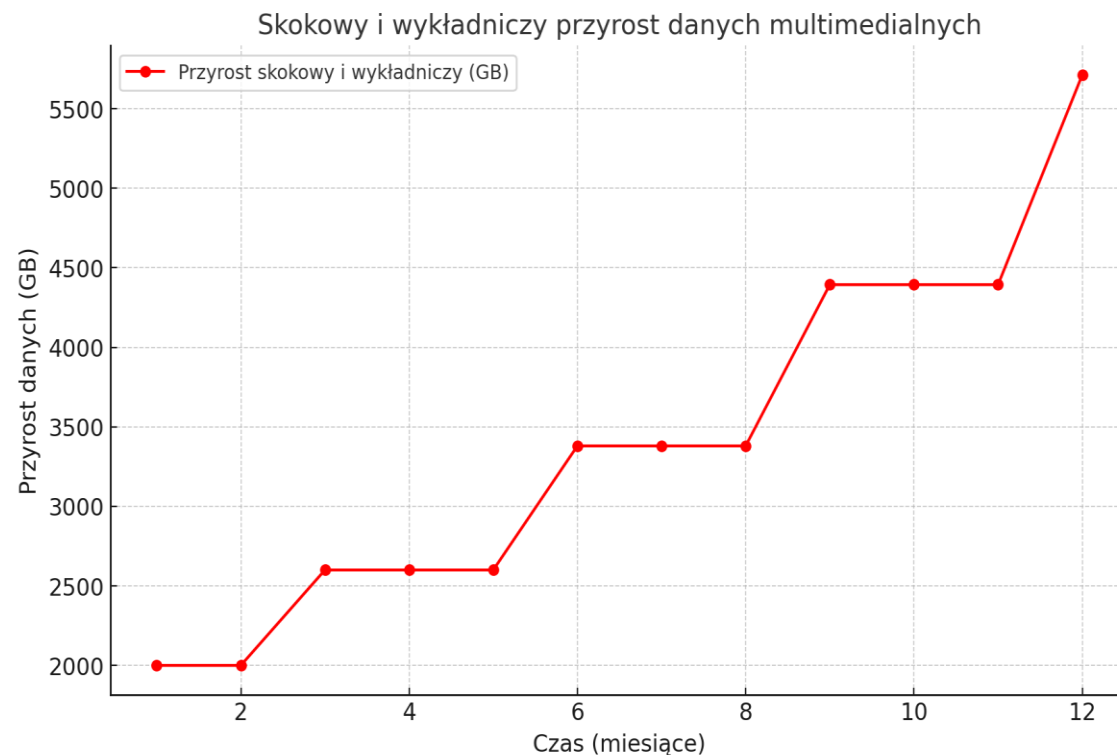
Gdzie:

P – całkowity przyrost danych w określonym przedziale czasu,

N – średnia liczba nowych plików multimedialnych w okresie,

D – liczba okresów (np. dni, miesiące),

S – średnia wielkość jednego nowego pliku multimedialnego (np.: GB)



Szacowany udział

40–50%

Dane geolokalizacyjne

Dane geolokalizacyjne (GIS)

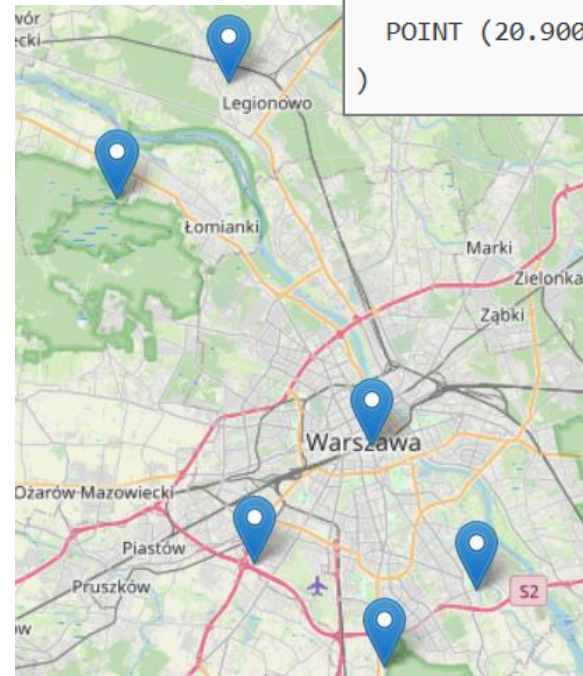
Dane przestrzenne, opisujące lokalizację obiektów w przestrzeni 2D lub 3D i ich atrybuty związane z lokalizacją

Przykład

Koordynaty GPS, mapy, warstwy GIS.

Typowe bazy / narzędzia

Rozszerzenia GIS w bazach relacyjnych
Bazy NoSQL z indeksami geograficznymi.



```
GEOMETRYCOLLECTION (  
  POINT (21.012230 52.229675 113), -- Warszawa  
  POINT (20.921112 52.173431 100), -- Pruszków  
  POINT (21.022915 52.123187 105), -- Piaseczno  
  POINT (20.811897 52.348571 120), -- Ożarów Mazowiecki  
  POINT (21.094776 52.160839 102), -- Konstancin-Jeziorna  
  POINT (20.900802 52.404239 118) -- Legionowo  
)
```

Dane geolokalizacyjne

Przyrost danych

Stopniowy wzrost (około 5-10% rocznie); intensywny w branżach, gdzie częste są nowe obiekty/lokalizacje (transport, logistyka, mapy).

$$P=O \times S \times D$$

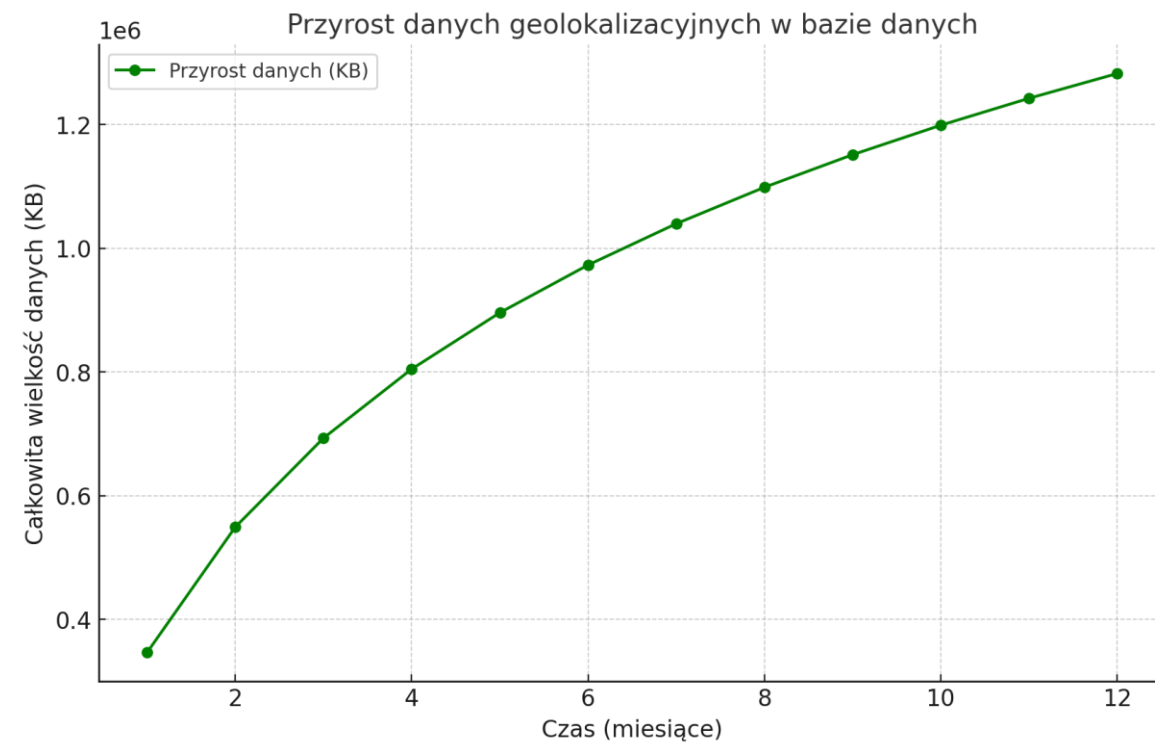
Gdzie:

P – całkowity przyrost danych w określonym przedziale czasu,

O – średnia liczba nowych obiektów w przestrzeni 2D lub 3D w okresie,

D – liczba okresów (np. dni, miesięcy),

S – średnia wielkość danych dla obiektów (np.: KB)



Szacowany udział

5–10%

Dane tymczasowe

Dane tymczasowe

Dane o krótkiej żywotności, przechowywane tymczasowo.

Przykład

Sesje użytkowników, tokeny autoryzacyjne, kolejki tymczasowych zadań, koszyki zakupowe lub cooki.

Typowe bazy / narzędzia

Bazy in-memory

Systemy kolejkowania i komunikatów

Bazy relacyjne

```
{
  "cart_id": "cart001",
  "user_id": "user123",
  "items": [
    {
      "product_id": "prod567",
      "product_name": "Smartphone XYZ",
      "quantity": 1,
      "price_per_unit": 2999.99,
      "total_price": 2999.99
    },
    {
      "product_id": "prod890",
      "product_name": "Wireless Charger",
      "quantity": 2,
      "price_per_unit": 199.99,
      "total_price": 399.98
    }
  ],
  "created_at": "2024-12-30T10:00:00Z",
  "expires_at": "2024-12-30T12:00:00Z"
}
```

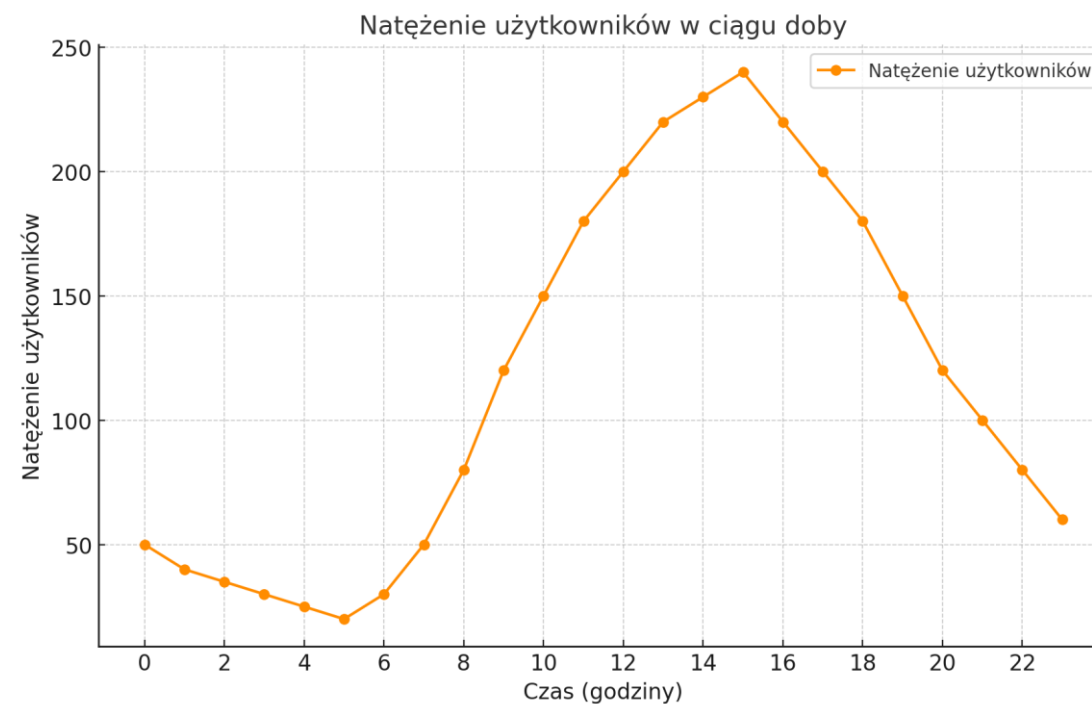


Set-Cookie: user_session=abc123xyz789; Domain=.example.com; Path=/; Secure; HttpOnly; SameSite=Strict; Expires=Tue, 30 Jan 2024 15:00:00 GMT

Dane tymczasowe

Przyrost danych

Zmienne i ulotne, trudno mówić o długoterminowym przyroście – często wygasają, zastępowane nowymi. Przyrosty są zmienne w ciągu dnia i zależą w dużej mierze od aktywności użytkowników



Szacowany udział

5–10%

Rodzaje danych

Przyrost danych w ciągu roku (12 miesięcy)

