

Temat 3 .

dr inż. Michał Malinowski

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

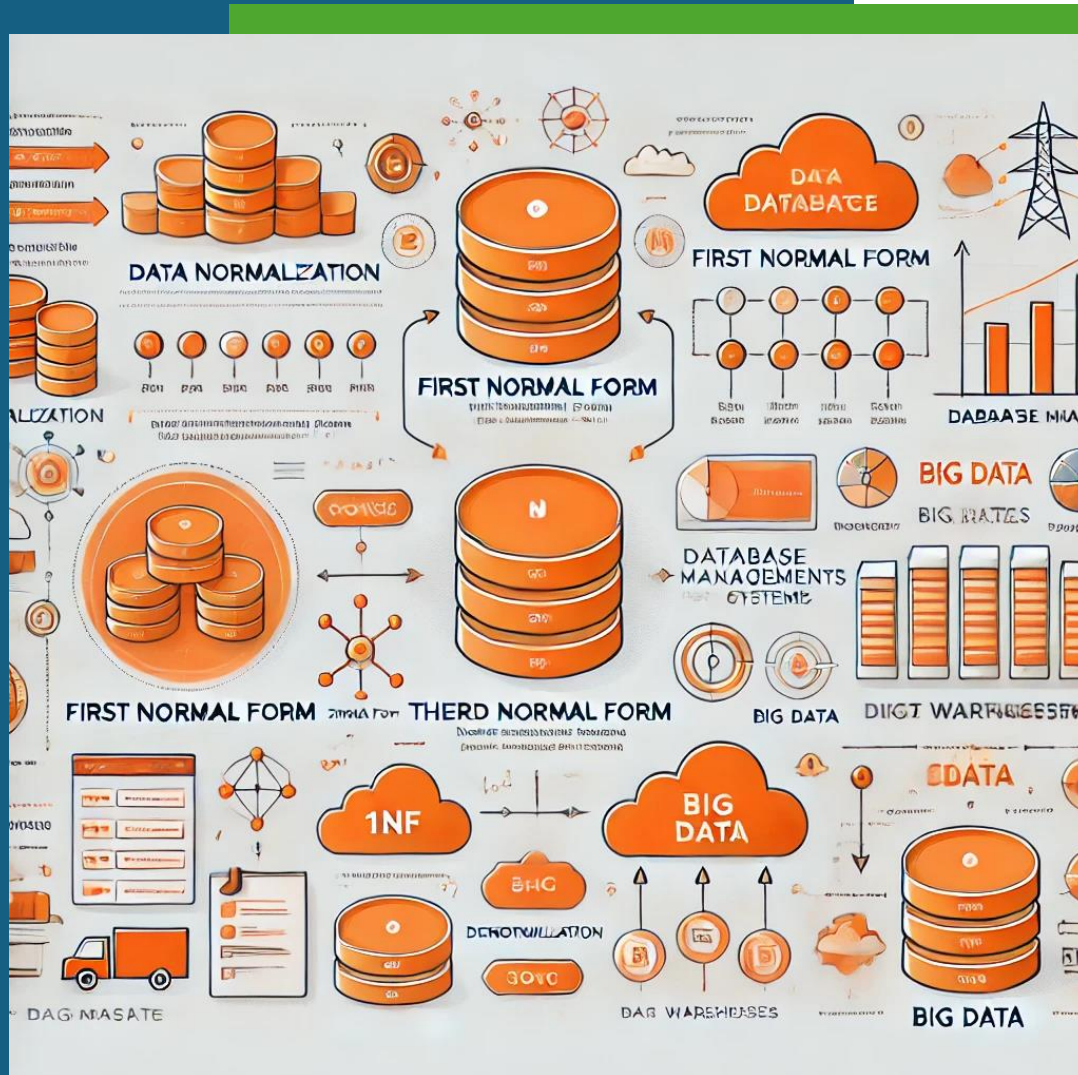
Normalizacja danych i duże bazy danych

PODSTAWY BAZ DANYCH



Zagadnienia:

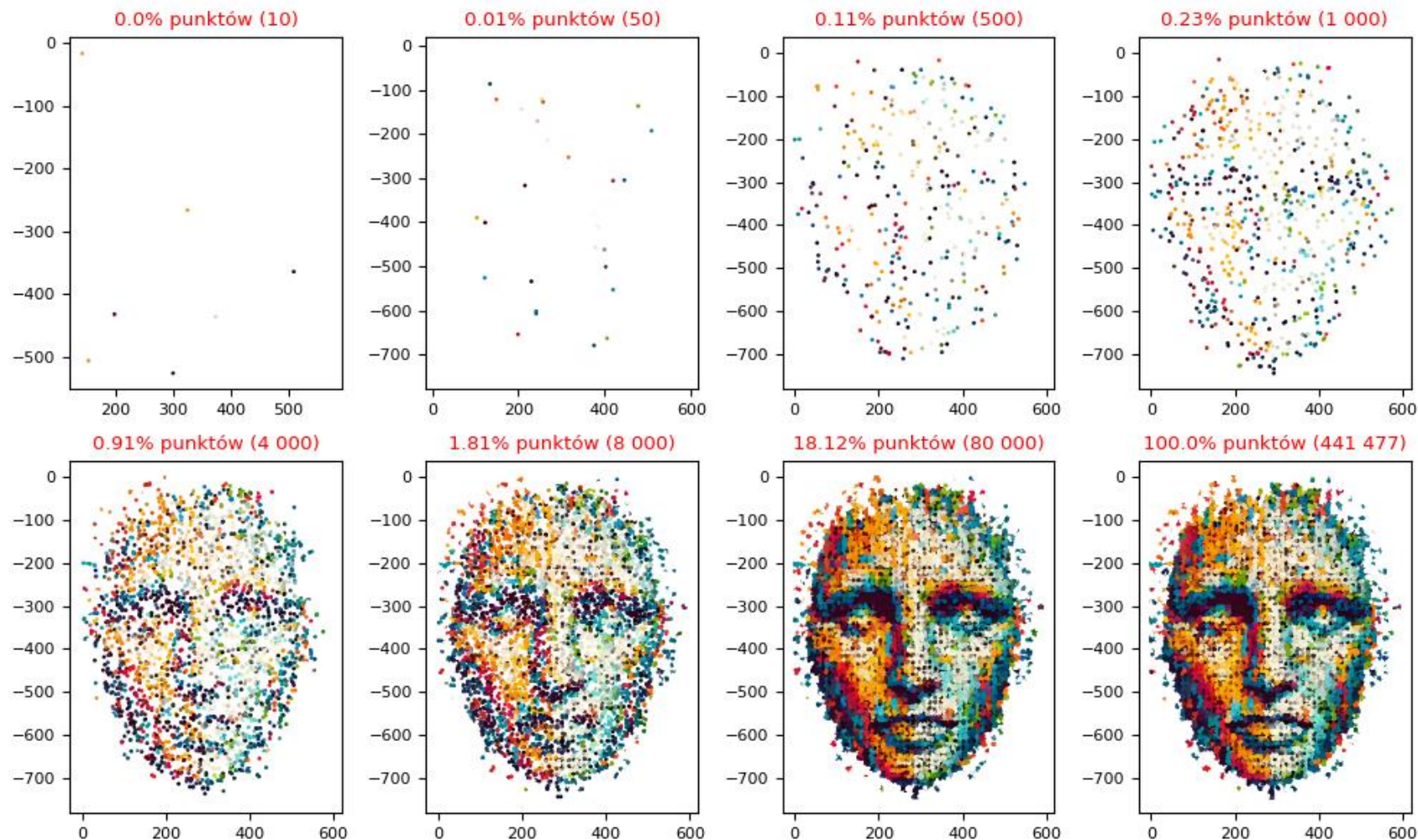
- Normalizacja danych
- Pierwsza postać normalna (1NF)
- Druga postać normalna (2NF)
- Trzecia postać normalna (3NF)
- De-normalizacja
- System Zarządzania Bazą Danych
- Hurtownie danych
- Big Data
- Usługi chmurowe



Ile potrzeba danych aby zrozumieć informację?



Dla obrazu wystarczy ~20% danych



Ile potrzeba danych aby zrozumieć informację?



Dla tekstu potrzeba ~90% danych

50% znaków

o ój no g i or ac jny w s t i h
ka h n z c mienił ajob z eko om cz y n ca m ś i . wp y ę
a w t ie a pe t iał ln i os od r ej p c i i za ą dz a , po
k nik cję i r t n . W g l o ś i k je r z n ę t od y
zn n w ro rodu n c to rzy is an a wdr ż n u
o snych h ol i . W ym k e ś i e i totn jest z z mi nie , w
aki posób IT j n zros p a czy ówne e ani y
eg w y u .

70% znaków

Rozw j t n lo i i fo mac jn ch w t tnic ekadach zna z c
zm enił j braz e o mic ny n cały świecie . IT wpłynę y
n w zystk e aspekty zia alno ci gos odar j , p odukc ar ą d ania , po
kom nik c j i ma ketin . W zczeg lno ci , kraje r z i iet o no ow ł nac ne
wzro t dukty no ś ci , któ e czę o prz pisywa e są r ą a iu o o esnych
te lo i . t ym kontek ś i to e jest ozu ni , w ja o ób IT
wpływ ją n zrost ospodarc y jakie są główne mec an my te p w .

90% znaków

Rozwój technologii in o ac n h w ostatnich dekadach zna ząco mieni
krajob z ekonomiczny na całym świecie . IT wpłynęły na wszystkie aspekt
działalność gospodarczej , d produkcji i arządzania , o omu i ację i
mar eting . W sz zególności , kraje rozwin i te dnot wał znaczne
w rosty p odukt ywności , które czę s przypisywane są wdr żaniu nowoczes ych
technologii . W tym kontekście istotne j t zrozumieni , w jaki sposób
IT wpływa a na wzrost gospoda czy i akie s główne me hanizmy tego pływu .

60% znaków

R z j te hn logii n rma y ych osta ch dekadac
zna ąco zmi nił k ajobra eko czny ły ś e ie . I y ę y n
w z s kie e dział al i osp d cz j , od du c i
z rzą d a i , o kom ni c i rketing . W zegó no ś ci , r j ro wi iet
odno wa y z ac n z ty r uk w o c re czę pr y isyw ne s
wdra ż an u wo z snych te n lo ii W m onte ś ci tn j z o mie w
ki os b IT w ywa a wzr st gosp da zy i a e są ł wne mec anizmy
tego p w .

80% znaków

zwój tech logii i ormacyj ch w sta n ch d a ach znacz co z ie ł
kraj o raz e ono icz y na ałym wi cie . IT w ły ęły na ystkie aspek y
z ała no ś ci gosp d cz , od produkcji i zarz dza ia , po komunikacj e i
marketing . W s zczególności , kraje roz in i ete od ot wały naczne wzr sty
pr d ktywności , które czę sto przyp sy a e a w d ą a iu nowoczes ych t chnol ii .
W t m on k ś cie sto ne je t zrozumien e , jak spo b I wpływ j a wzros
gospodarc y j kie są głów echanizm tego wpływu

100% znaków

Rozwój technologii informacyjnych w ostatnich dekadach znacząco zmienił
krajobraz ekonomiczny na całym świecie . IT wpłynęły na wszystkie aspekty
działalności gospodarczej , od produkcji i zarządzania , po komunikację i
marketing . W szczególności , kraje rozwinięte odnotowały znaczne wzrosty
produktywności , które często przypisywane są wdrażaniu nowoczesnych technologii .
W tym kontekście istotne jest zrozumienie , w jaki sposób IT wpływają na wzrost
gospodarczy i jakie są główne mechanizmy tego wpływu .

Normalizacja danych

Redundancja

to nadmiarowość danych, która polega na powtarzaniu tych samych informacji w wielu miejscach w bazie danych.

Może prowadzić do zwiększenia rozmiaru bazy oraz problemów z integralnością, takich jak trudności w utrzymaniu spójnych danych podczas aktualizacji, wstawiania lub usuwania rekordów.

Eliminacja redundancji jest jednym z celów normalizacji, której celem jest zorganizowanie danych w sposób efektywny i spójny.

Redundancja pamięci:

- Powielanie tych samych danych zajmuje więcej przestrzeni dyskowej.

Anomalie aktualizacji:

- Zmiana wartości w jednym miejscu wymaga wprowadzenia zmian we wszystkich powielonych miejscach.

Anomalie usuwania:

- Usunięcie wartości wymaga jej usunięcia w wielu miejscach w bazie danych

Normalizacja danych

Problemy związane z redundancją

Redundancja pamięci

ID Klient	Imię	Miasto
1	Jan	Warszawa
2	Anna	Warszawa
3	Piotr	Warszawa

Wartość „Warszawa” powtarza się w kolumnie „Miasto” dla wielu klientów, co prowadzi do nadmiaru danych w tabeli.

Anomalie aktualizacji

ID Oferta	Produkt	Cena
101	Laptop VX	3000
102	Laptop VX	3000
103	Laptop VX	3000

Zmiana ceny produktu „Laptop VX” wymaga zaktualizowania wartości we wszystkich ofertach, gdzie produkt ten występuje.

Anomalie usuwania

ID Klient	Imię	ID Zamówienie	Produkt
1	Jan	101	Laptop VX
2	Anna	102	Tablet CC

Usunięcie zamówienia numer 101 prowadzi do utraty informacji o kliencie Janie, ponieważ dane klienta są przechowywane razem z zamówieniem.

Normalizacja danych

Normalizacja

to proces organizacji danych w bazie w taki sposób, aby zminimalizować redundancję i wyeliminować niepożądane anomalie aktualizacji, usuwania i wstawiania.

Celem normalizacji jest stworzenie struktury danych, która zapewnia spójność, efektywność oraz integralność danych.

Redukcja redundancji danych:

- Eliminowanie powtarzających się danych, co prowadzi do mniejszego zużycia przestrzeni dyskowej i łatwiejszego zarządzania danymi.

Zwiększenie spójności danych:

- Upewnienie się, że dane w bazie są zgodne, a różne zapytania lub operacje na danych zwracają poprawne i spójne wyniki.

Poprawa integralności danych:

- Wprowadzenie mechanizmów zapewniających, że dane są poprawne i zgodne z regułami biznesowymi, np. poprzez stosowanie kluczy głównych i obcych.

Ułatwienie modyfikacji struktury bazy danych:

- Struktura bazy jest bardziej elastyczna i łatwiejsza do dostosowania do nowych wymagań bez konieczności zmiany wielu części bazy danych.

Pierwsza postać normalna (1NF)

Pierwsza postać normalna (1NF)

to podstawowy poziom normalizacji danych w bazach danych, który wymaga, aby tabela spełniała następujące warunki:

1. Każda kolumna zawiera wartości atomowe:

Kolumny nie mogą zawierać złożonych danych, takich jak listy, tablice lub wielokrotne wartości w jednym polu. Każda kolumna powinna przechowywać pojedynczą wartość.

Przykład: W tabeli dotyczącej zamówień, zamiast kolumny „Produkty” z wartościami ["Laptop", "Telefon"], powinniśmy mieć oddzielny wiersz dla każdego produktu.

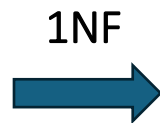
Pierwsza postać normalna (1NF)

2. Brak powtarzających się grup danych:

Tabela nie powinna zawierać kolumn, które przechowują wiele wartości tego samego typu dla jednego rekordu.

Przykład: Jeśli w jednej tabeli zapisujemy zamówienia, każde zamówienie powinno mieć własny wiersz, a nie jedno zamówienie z wieloma produktami w jednej kolumnie.

ZamówienieID	Klient	Produkty
101	Jan Nowak	Laptop, Telefon
102	Anna Kowalska	Tablet, Smartwatch



ZamówienieID	Imię	Nazwisko	Produkt
101	Jan	Nowak	Laptop
101	Jan	Nowak	Telefon
102	Anna	Kowalska	Tablet
102	Anna	Kowalska	Smartwatch

Druga postać normalna (2NF)

Druga postać normalna (2NF)

to drugi poziom normalizacji danych w bazach danych, który wymaga, aby tabela spełniała następujące warunki:

1. Była spełniona 1 postać normalna (1NF)

Druga postać normalna (2NF)

2. Eliminacja częściowych zależności:

Wszystkie atrybuty, które nie są kluczem głównym, muszą być w pełni zależne od całego klucza głównego, a nie tylko od jego części.

Zależności częściowe występują, gdy kolumna jest zależna tylko od fragmentu klucza głównego w przypadku złożonego klucza głównego (składającego się z więcej niż jednej kolumny).

ZamówienieID	ProduktID	Imię	Nazwisko	Adres	Produkt	Cena
101	1	Jan	Nowak	Warszawa	Laptop	3000
102	2	Anna	Kowalska	Gdańsk	Telefon	1500

2NF



Tabela zostaje rozbita na dwie oddzielne tabele, eliminując częściowe zależności.

ZamówienieID	Imię	Nazwisko	Adres
101	Jan	Nowak	Warszawa
102	Anna	Kowalska	Gdańsk

ZamówienieID	ProduktID	Produkt	Cena
101	1	Laptop	3000
102	2	Telefon	1500

Trzecia postać normalna (3NF)

Trzecia postać normalna (3NF)

to trzeci poziom normalizacji danych w bazach danych, który wymaga, aby tabela spełniała następujące warunki:

1. Była spełniona 2 postać normalna (2NF)
2. Każda kolumna niebędąca kluczem głównym jest bezpośrednio zależna od klucza głównego, a nie od innej kolumny w tabeli (brak zależności przechodnich).
Oznacza to, że nie może być tak, że jedna kolumna zależy od innej, a ta inna od klucza głównego.

Trzecia postać normalna (3NF)

2. Każda kolumna niebędąca kluczem głównym jest bezpośrednio zależna od klucza głównego, a nie od innej kolumny w tabeli (brak zależności przechodnich).
Oznacza to, że nie może być tak, że jedna kolumna zależy od innej, a ta inna od klucza głównego.

ZamówienieID	KlientID	Imię	Nazwisko	Miasto	Sprzedawca	Sprzedawca Telefon
101	1	Jan	Nowak	Warszawa	Jan	123-456-789
102	2	Anna	Kowalska	Gdańsk	Jan	123-456-789

3NF



ZamówienieID	KlientID	Imię	Nazwisko	Miasto	Sprzedawca
101	1	Jan	Nowak	Warszawa	Jan
102	2	Anna	Kowalska	Gdańsk	Jan

Sprzedawca	Telefon
Jan	123-456-789

W tym przykładzie klucz główny to ZamówienieID, ale kolumna SprzedawcaTelefon zależy od Sprzedawca, a nie bezpośrednio od ZamówienieID.

Jest to zależność przechodnia, ponieważ SprzedawcaTelefon zależy pośrednio od ZamówienieID za pośrednictwem Sprzedawca

W tabeli "Zamówienia" każda kolumna zależy bezpośrednio od klucza głównego ZamówienieID,
a tabela "Sprzedawcy" przechowuje dane o sprzedawcach.
W ten sposób eliminujemy zależność przechodnią pomiędzy SprzedawcaTelefon a ZamówienieID, która wcześniej występowała za pośrednictwem kolumny Sprzedawca.

Trzecia postać normalna (3NF)

ZamówienieID	Klient	Miasto	Produkt	Cena	Sprzedawca	Telefon
101	Jan Nowak	Warszawa	Laptop	3000	Jan	123-456-789
102	Anna Kowalska	Gdańsk	Telefon	1500	Jan	123-456-789

Normalizacja
(1NF, 2NF i 3NF)



De-normalizacja

ProduktID	Produkt	Cena
1	Laptop	3000
2	Telefon	1500

KlientID	Imię	Nazwisko	Miasto
1	Jan	Nowak	Warszawa
2	Anna	Kowalska	Gdańsk

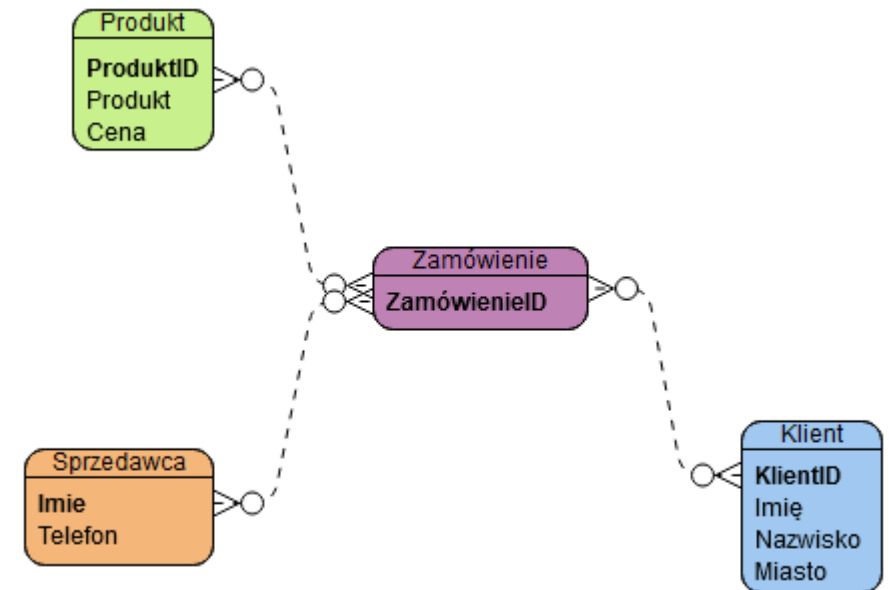
Sprzedawca	Telefon
Jan	123-456-789

ZamówienieID	ProduktID
101	1
102	2

ZamówienieID	KlientID
101	1
102	2

ZamówienieID	Sprzedawca
101	Jan
102	Jan

Diagram ERD



De-normalizacja

De-normalizacja

to proces odwrotny do normalizacji, w którym celowo wprowadza się redundancję danych w bazie danych, aby poprawić wydajność zapytań.

Podczas gdy normalizacja minimalizuje redundancję i optymalizuje spójność danych, de-normalizacja polega na wprowadzeniu nadmiarowych danych w celu redukcji liczby operacji złączeń (JOIN) między tabelami.

Wydajność odczytu jest priorytetem:

- Gdy baza danych obsługuje bardzo wiele operacji odczytu (np. systemy hurtowni danych, aplikacje z dużą liczbą zapytań odczytujących), de-normalizacja może być korzystna.

Systemy raportowania:

- De-normalizacja jest powszechnie stosowana w systemach analitycznych i hurtowniach danych, gdzie czas odpowiedzi na zapytania jest kluczowy.

Koszt operacji JOIN jest wysoki:

- W sytuacjach, gdy wykonywanie złączeń na dużych tabelach drastycznie obniża wydajność, de-normalizacja może przyspieszyć dostęp do danych.

De-normalizacja

Wady de-normalizacji

Zwiększona redundancja:

- De-normalizacja powoduje powielanie danych, co może prowadzić do zwiększenia rozmiaru bazy danych.

Trudniejsze utrzymanie spójności:

- Aktualizowanie danych w zde-normalizowanej bazie staje się bardziej skomplikowane, ponieważ ta sama informacja może występować w wielu miejscach. Może to prowadzić do problemów z integralnością danych, jeśli dane nie zostaną poprawnie zsynchronizowane.

Trudniejsze zarządzanie:

- Zwiększona redundancja powoduje konieczność bardziej złożonego zarządzania danymi i utrzymania ich zgodności, zwłaszcza przy operacjach wstawiania, aktualizacji i usuwania danych.

System Zarządzania Bazą Danych

SZBD - System Zarządzania Bazą Danych (DBMS, Database Management System)

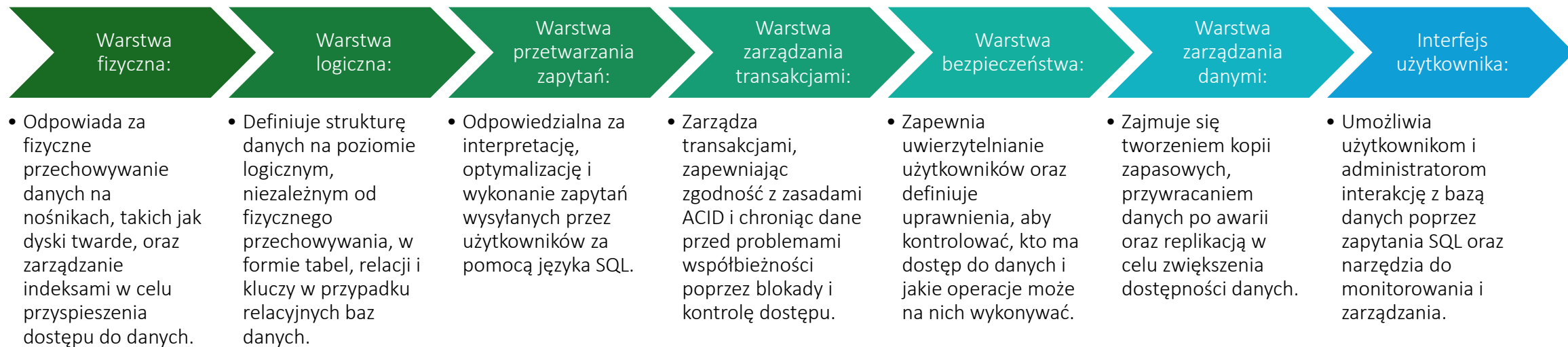
Oprogramowanie umożliwiające tworzenie, zarządzanie, manipulowanie i kontrolowanie dostępu do bazy danych.

Funkcje



System Zarządzania Bazą Danych

Architektura

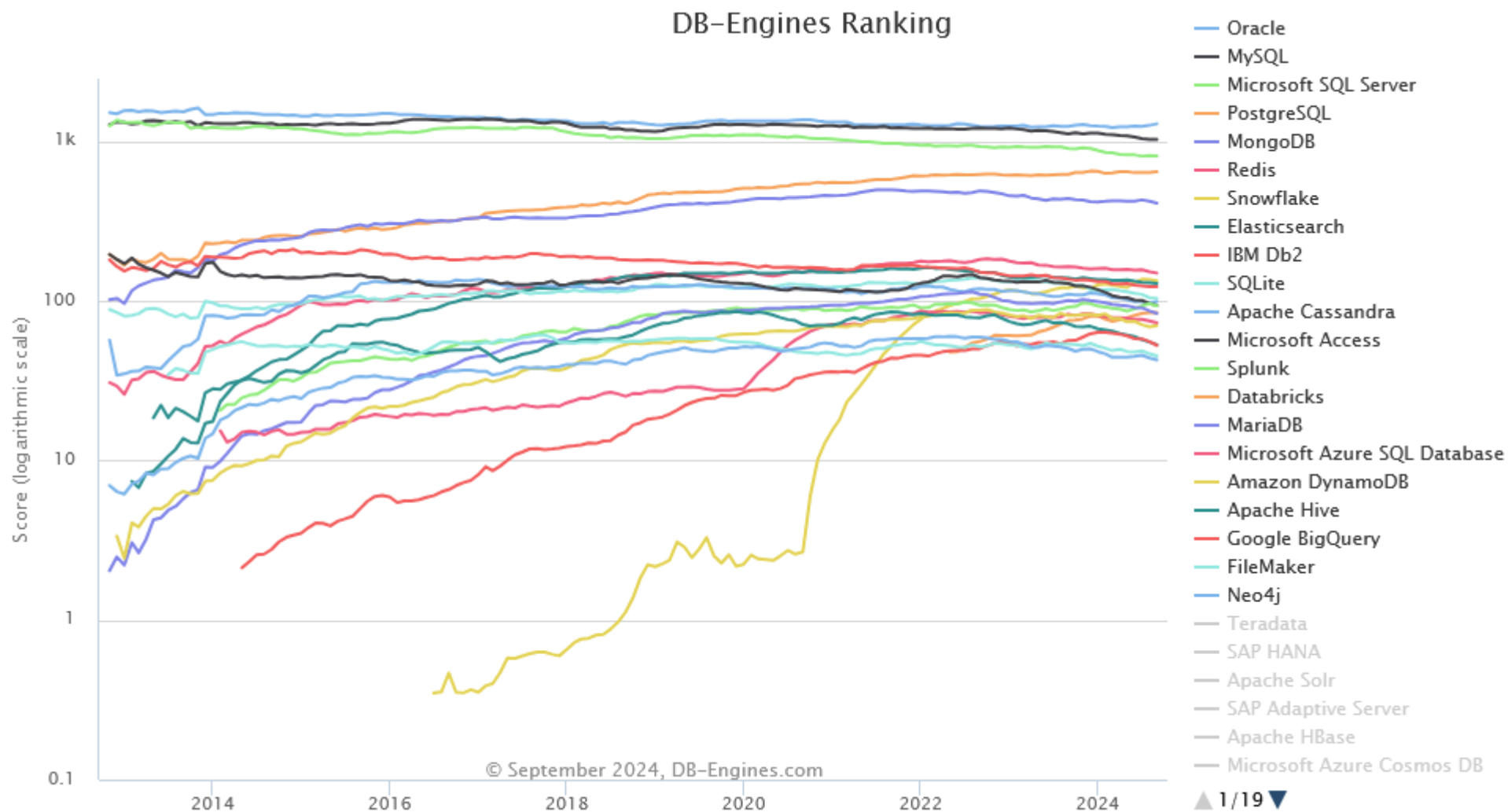


Ranking - Systemy baz danych

423 systems in ranking, September 2024

Rank			DBMS	Database Model	Score		
Sep 2024	Aug 2024	Sep 2023			Sep 2024	Aug 2024	Sep 2023
1.	1.	1.	Oracle +	Relational, Multi-model i	1286.59	+28.11	+45.72
2.	2.	2.	MySQL +	Relational, Multi-model i	1029.49	+2.63	-82.00
3.	3.	3.	Microsoft SQL Server +	Relational, Multi-model i	807.76	-7.41	-94.45
4.	4.	4.	PostgreSQL +	Relational, Multi-model i	644.36	+6.97	+23.61
5.	5.	5.	MongoDB +	Document, Multi-model i	410.24	-10.74	-29.18
6.	6.	6.	Redis +	Key-value, Multi-model i	149.43	-3.28	-14.26
7.	7.	↑ 11.	Snowflake +	Relational	133.72	-2.25	+12.83
8.	8.	↓ 7.	Elasticsearch	Search engine, Multi-model i	128.79	-1.04	-10.20
9.	9.	↓ 8.	IBM Db2	Relational, Multi-model i	123.05	+0.04	-13.67
10.	10.	↓ 9.	SQLite +	Relational	103.35	-1.44	-25.85
11.	11.	↑ 12.	Apache Cassandra +	Wide column, Multi-model i	98.94	+1.94	-11.11
12.	12.	↓ 10.	Microsoft Access	Relational	93.76	-2.61	-34.81
13.	13.	↑ 14.	Splunk	Search engine	93.02	-3.08	+1.63
14.	↑ 15.	↑ 17.	Databricks +	Multi-model i	84.24	-0.22	+9.06
15.	↓ 14.	↓ 13.	MariaDB +	Relational, Multi-model i	83.44	-3.09	-17.01
16.	16.	↓ 15.	Microsoft Azure SQL Database	Relational, Multi-model i	72.95	-2.08	-9.78
17.	17.	↓ 16.	Amazon DynamoDB +	Multi-model i	70.06	+1.15	-10.85
18.	↑ 19.	18.	Apache Hive	Relational	53.07	-2.17	-18.76
19.	↓ 18.	↑ 20.	Google BigQuery +	Relational	52.67	-2.86	-3.80
20.	20.	↑ 21.	FileMaker	Relational	45.20	-1.47	-8.40
21.	21.	↑ 23.	Neo4j +	Graph	42.68	-1.22	-7.71
22.	↑ 23.	↓ 19.	Teradata	Relational, Multi-model i	41.47	-0.78	-18.86
23.	↓ 22.	↓ 22.	SAP HANA +	Relational, Multi-model i	41.32	-1.04	-9.29
24.	24.	24.	Apache Solr	Search engine, Multi-model i	34.15	-2.14	-12.94
25.	25.	25.	SAP Adaptive Server	Relational, Multi-model i	32.67	-0.61	-10.64

Ranking - Systemy baz danych



Hurtownie danych

Hurtownia danych

to centralny, zintegrowany system do przechowywania dużych ilości danych pochodzących z różnych źródeł.

Jest wykorzystywany głównie do celów analitycznych i raportowania, umożliwiając organizacjom analizowanie historycznych i bieżących danych, wspierając podejmowanie decyzji na poziomie strategicznym.

Cechy

Zintegrowane źródło danych:

Konsoliduje informacje z różnych systemów (np. systemy CRM, ERP, e-commerce) w jednym miejscu, zapewniając jednolitą strukturę danych.

Tematyczne przechowywanie danych:

Dane są organizowane wokół głównych tematów biznesowych, takich jak sprzedaż, finanse, logistyka czy zasoby ludzkie.

Długoterminowe przechowywanie danych:

Przechowują dane historyczne przez dłuższy czas, umożliwiając analizowanie trendów i wzorców w oparciu o dane sprzed wielu lat.

Stołość danych:

Są zwykle niezmiennie. Po załadowaniu do hurtowni rzadko się je aktualizuje..

Dane zorientowane na użytkownika:

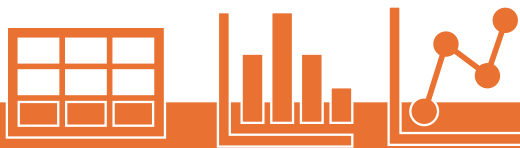
Jest optymalizowana pod kątem zapytań analitycznych (De-normalizacja).

Hurtownie danych

Architektura

Warstwa źródłowa (Źródła danych):

- Różnorodne systemy, z których pobierane są dane, np. bazy danych, systemy transakcyjne, pliki CSV, dane z API, logi systemowe. Dane są pozyskiwane z wielu formatów i źródeł.



Proces ETL (Extract, Transform, Load):

- Kluczowy element hurtowni danych, w którym dane są przetwarzane:



Warstwa hurtowni danych:

- Główna baza danych, w której przechowywane są dane, zoptymalizowana do odczytu. Dane są zazwyczaj organizowane w formie schematów gwiazdy lub płatka śniegu, co umożliwia szybkie zapytania analityczne.



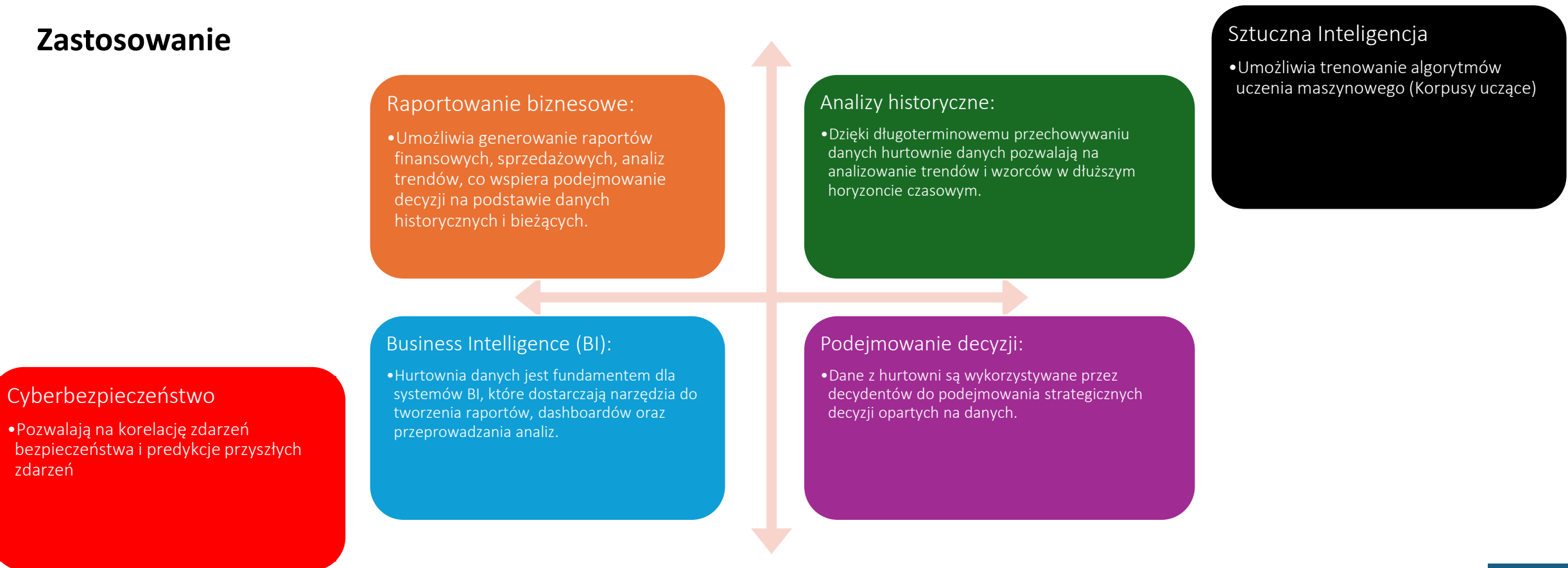
Warstwa dostępu do danych:

- Użytkownicy końcowi, analitycy, systemy raportowania i narzędzia BI (Business Intelligence) uzyskują dostęp do danych z hurtowni za pośrednictwem interfejsów zapytań, raportów lub dashboardów.



Hurtownie danych

Zastosowanie



Jednostki wielkości pamięci i mocy procesora

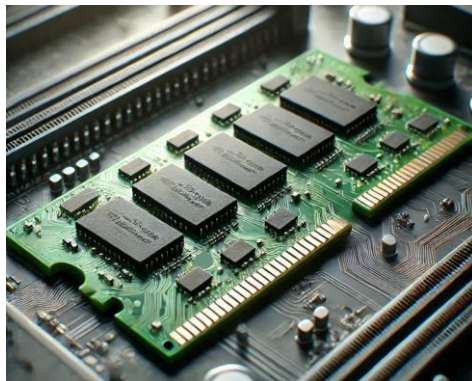
Moc procesora

Jednostka	Przelicznik	Liczba cykli (s)	Przykłady rzeczywistych wielkości
Cykl zegara	1 cykl	1	Najmniejsza jednostka pracy procesora
Kiloherc (KHz)	1 tysiąc cykli na sekundę	1 000	Wczesne mikroprocesory z lat 70 i 80
Megaherc (MHz)	1 milion cykli na sekundę	1 000 000	Taktowanie procesorów z lat 90.
Gigaherc (GHz)	1 miliard cykli na sekundę	1 000 000 000	Obecne procesory pracujące z częstotliwościami w GHz
Teraherc (THz)	1 bilion cykli na sekundę	1 000 000 000 000	Teoretyczna wartość, aktualnie nieosiągalna

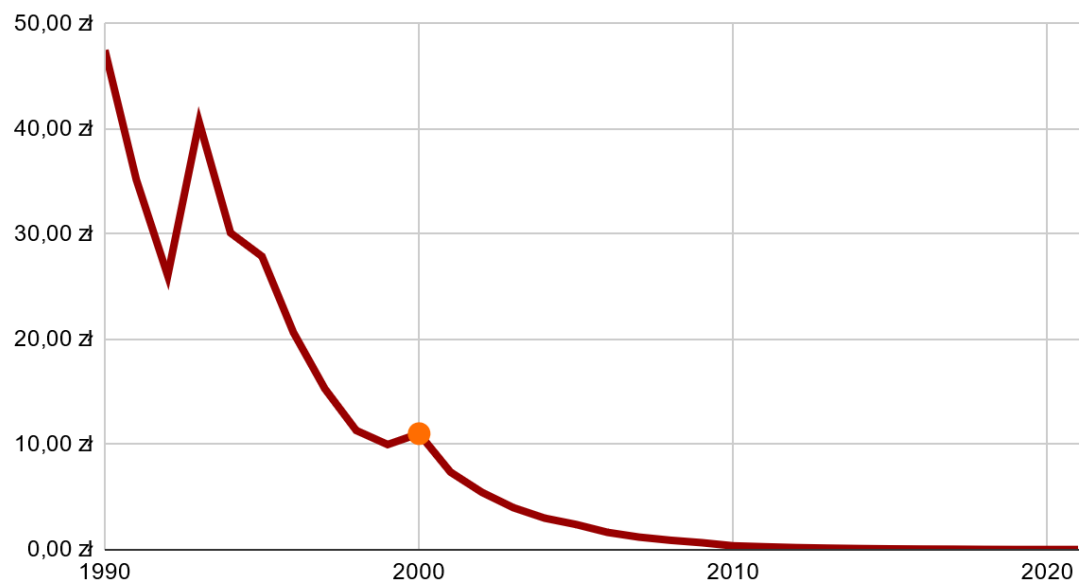
Wielkości pamięci

Jednostka	Przelicznik	Liczba bajtów	Dziesiętnie		Przykłady rzeczywistych wielkości
bit (1)	1 bit				Najmniejsza jednostka, przyjmuje wartości 0 lub 1
Bajt (B)	8 bitów	1	1	jeden	Jeden znak w tekście przyjmuje wartości od 0 do 255 (kod znaku)
Kilobajt (KB)	1024 B	1 024	10^3	tysiąc	Mały plik tekstowy lub niewielki obrazek JPEG
Megabajt (MB)	1024^2 B	1 048 576	10^6	milion	Piosenka w formacie MP3 lub krótki filmik w jakości SD
Gigabajt (GB)	1024^3 B	1 073 741 824	10^9	miliard	System operacyjny, filmy w jakości HD lub gry komputerowe
Terabajt (TB)	1024^4 B	1 099 511 627 776	10^{12}	bilion	Duża baza danych, kolekcja filmów lub zasoby dużego przedsiębiorstwa
Petabajt (PB)	1024^5 B	1 125 899 906 842 620	10^{15}	bilion	Zbiory danych w dużych serwerowniach, archiwa bibliotek cyfrowych
Eksabajt (EB)	1024^6 B	1 152 921 504 606 850 000	10^{18}	trylion	Całkowita ilość danych generowana przez Internet w ciągu kilku dni
Zettabajt (ZB)	1024^7 B	1 180 591 620 717 410 000 000	10^{21}	tryliard	Oszacowana ilość wszystkich danych istniejących w cyberprzestrzeni

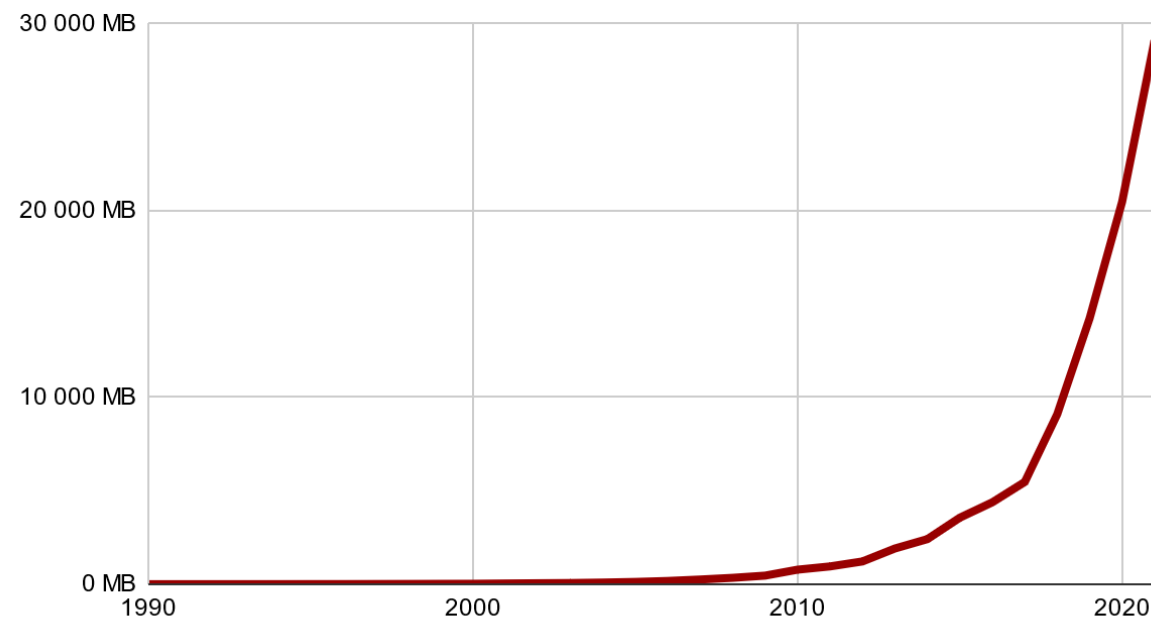
Średnia cena pamięci RAM



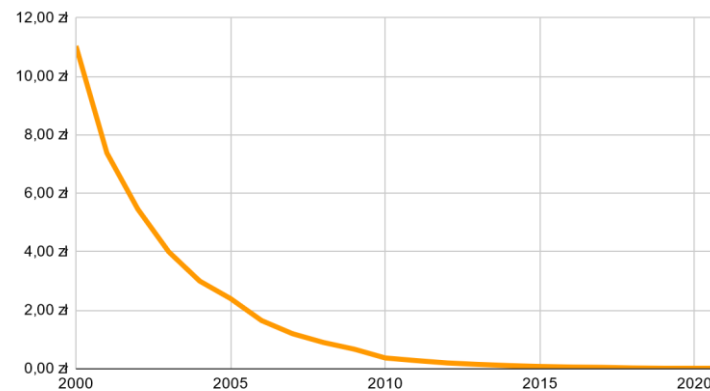
Średnia cena pamięci RAM (1 MB, komputer osobisty)



Średnie wykorzystanie RAM na komputer osobisty



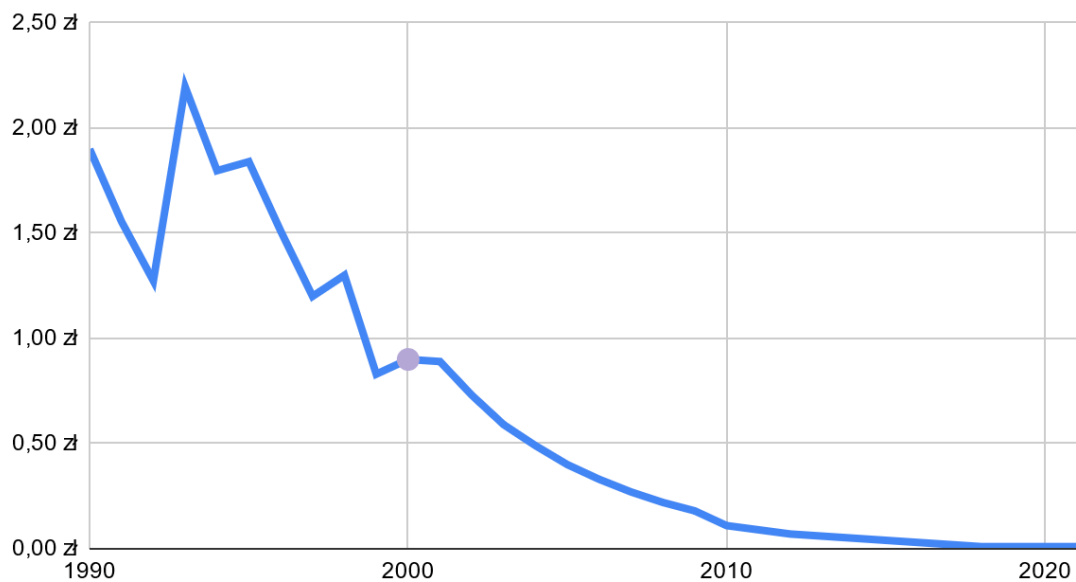
Średnia cena pamięci RAM (1 MB, komputer osobisty)



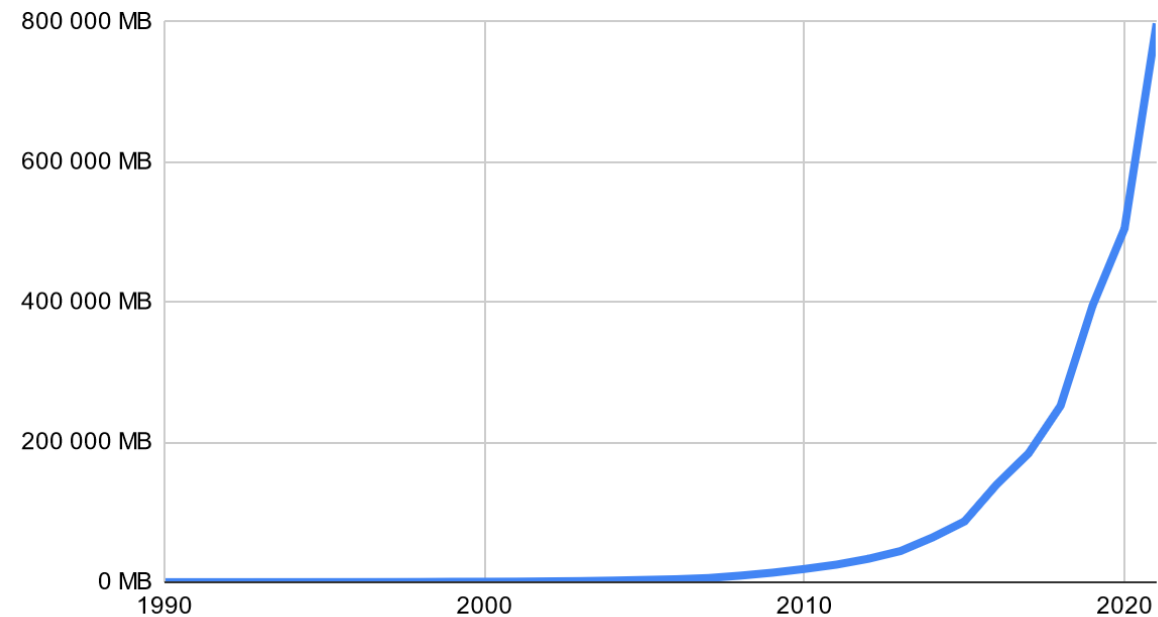
Średnia cena pamięci HDD



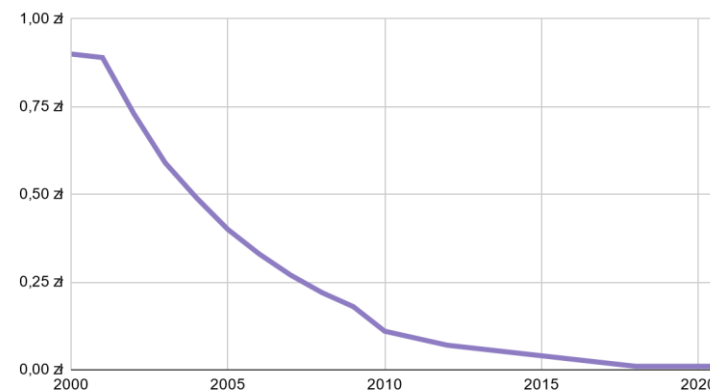
Średnia cena pamięci HDD (1 MB, komputer osobisty)



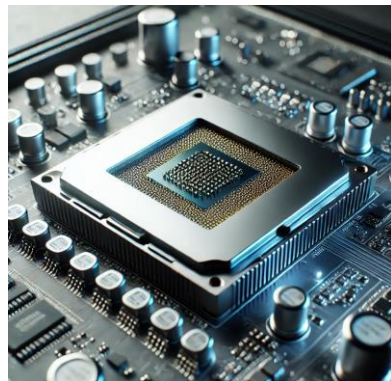
Średnie wykorzystanie HDD na komputer osobisty



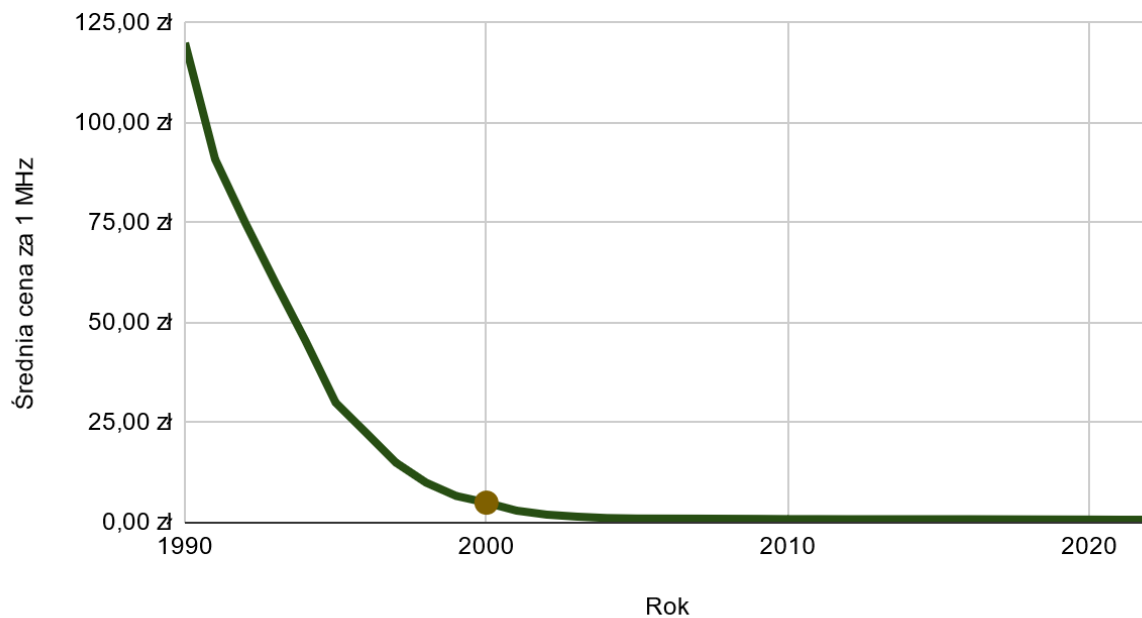
Średnia cena pamięci HDD (1 MB, komputer osobisty)



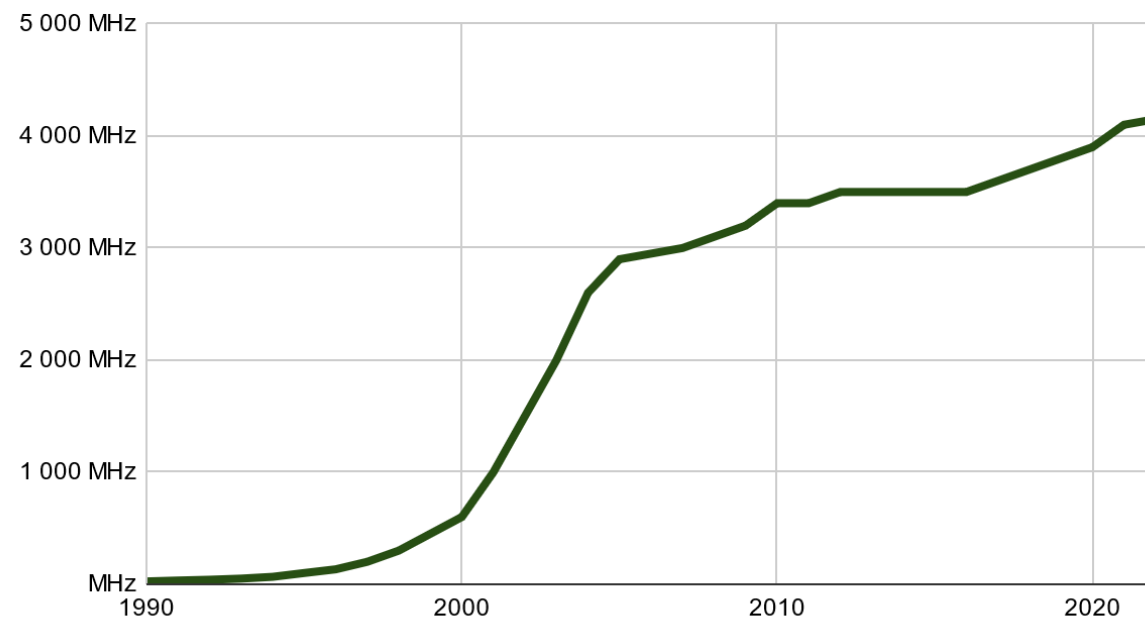
Średnia cena procesora



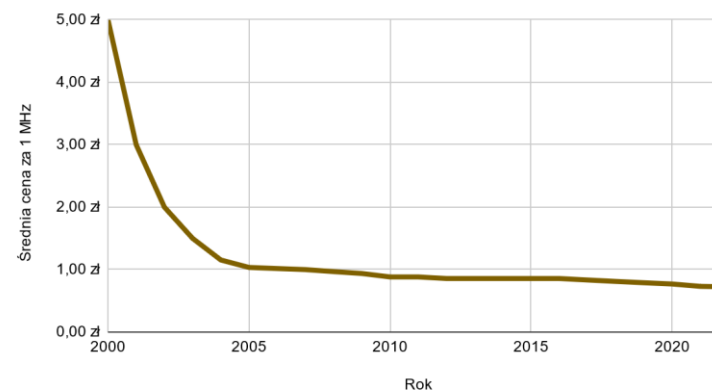
Średnia cena procesora (1 MHz, komputer osobisty)



Średnia częstotliwość procesora



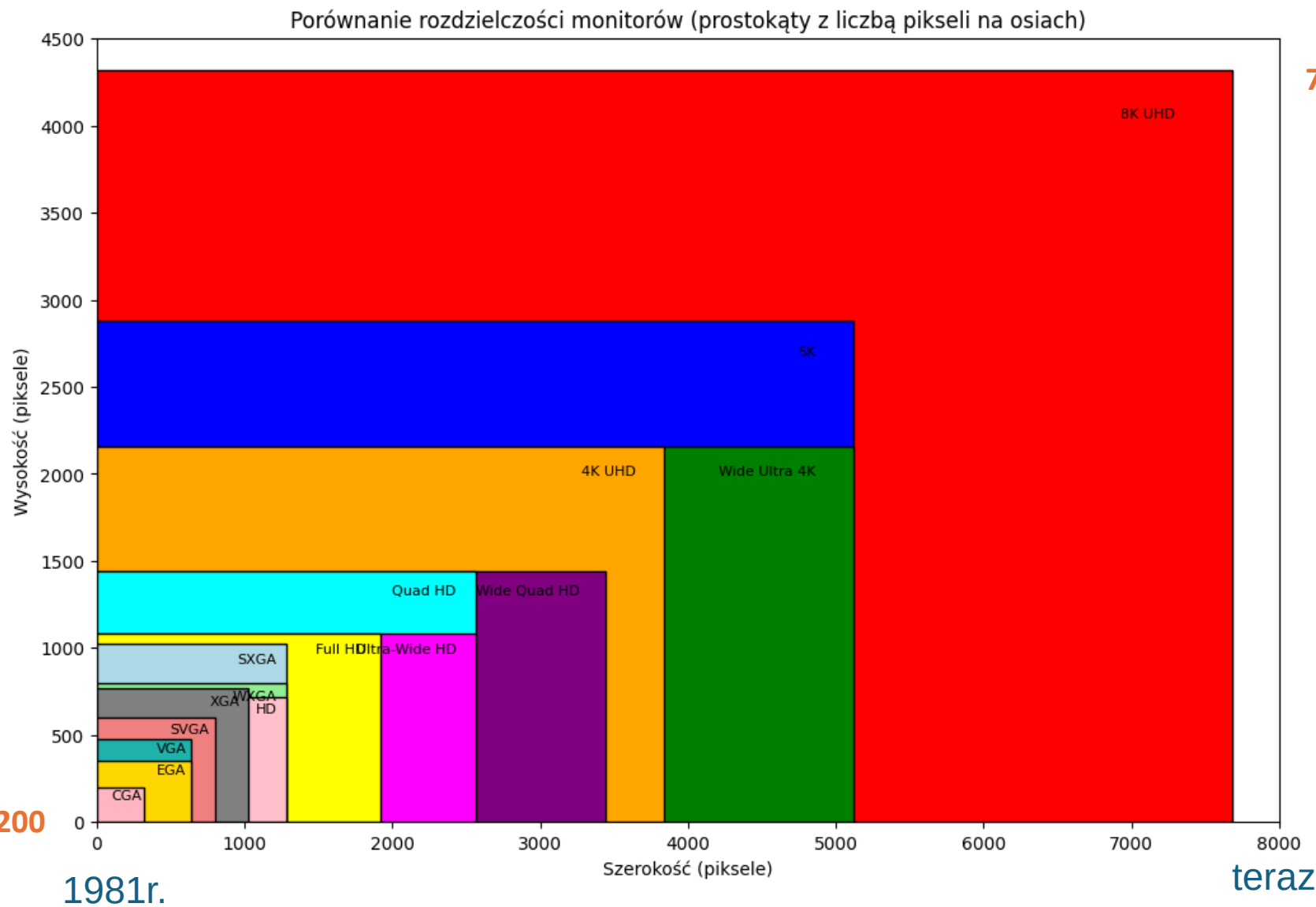
Średnia cena procesora (1 MHz, komputer osobisty)



Monitory (rozdzielczości)



320 x 200



7 680 x 4 320

Hurtownie danych

ETL vs ELT

ETL (Extract, Transform, Load)

ETL to tradycyjny proces przetwarzania danych, który składa się z trzech głównych etapów:

1. **Extract (Ekstrakcja):** Dane są pobierane ze źródeł, takich jak bazy danych, pliki czy aplikacje.
2. **Transform (Transformacja):** Dane są przekształcane, czyszczone, łączone i formatowane zgodnie z wymaganiami docelowej bazy danych lub aplikacji.
3. **Load (Ładowanie):** Przekształcone dane są ładowane do docelowej hurtowni danych lub bazy danych.

ELT (Extract, Load, Transform)

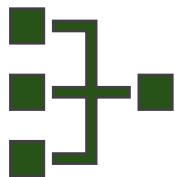
ELT to nowsze podejście do przetwarzania danych, które również składa się z trzech etapów, ale różni się kolejnością:

1. **Extract (Ekstrakcja):** Dane są pobierane ze źródeł, podobnie jak w ETL.
2. **Load (Ładowanie):** Surowe dane są ładowane bezpośrednio do docelowej hurtowni danych lub bazy danych.
3. **Transform (Transformacja):** Transformacja danych odbywa się po ich załadowaniu, zazwyczaj przy użyciu mocy obliczeniowej samej hurtowni danych lub systemu analitycznego.

Hurtownie danych

Źródła surowych danych

ETL



Extract



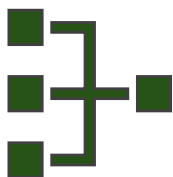
Transform

Load



Docelowa baza danych

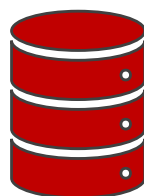
ELT



Extract



Load



Transform

Wybierz ETL, gdy:

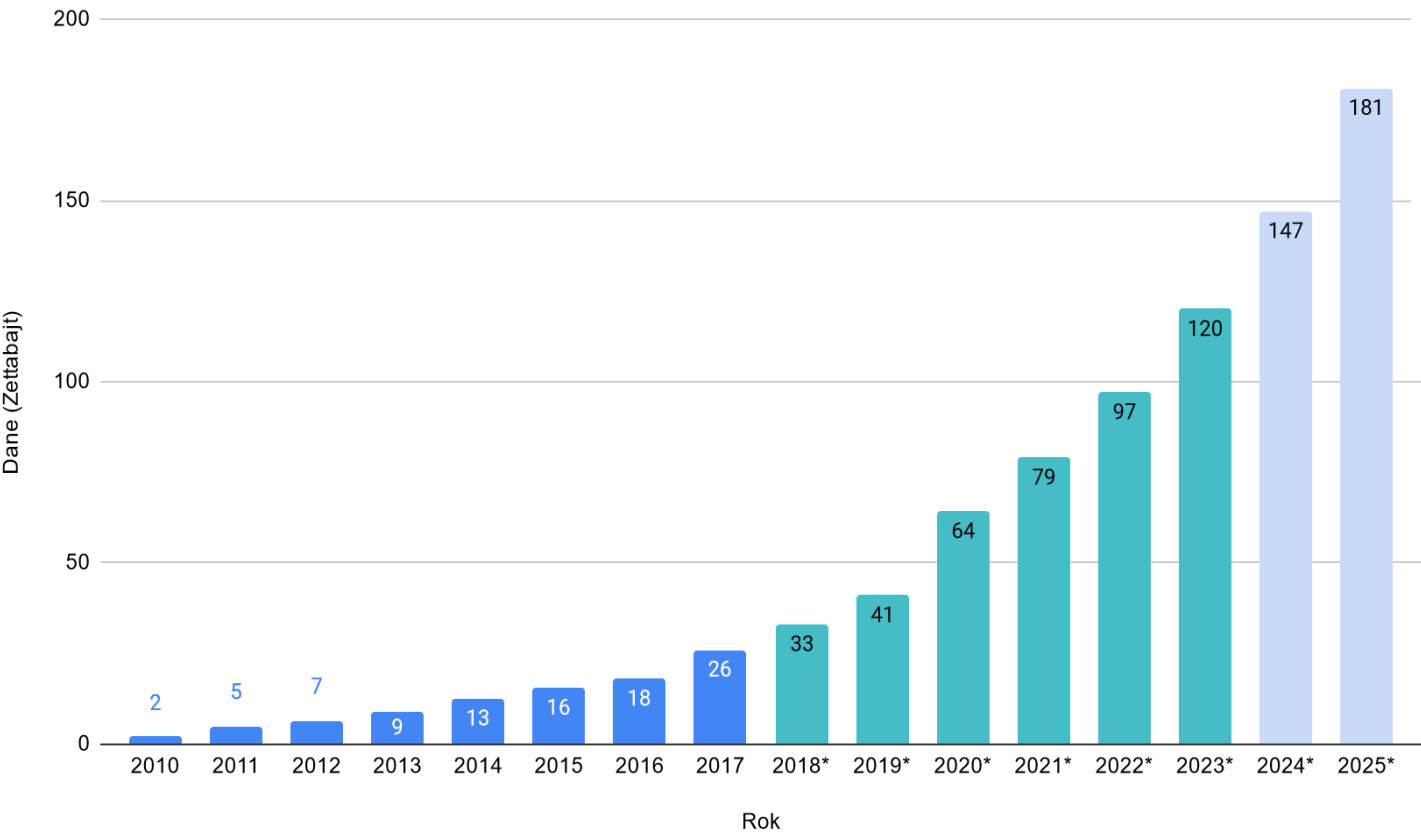
- Pracujesz z tradycyjnymi systemami hurtowni danych.
- Jakość i integralność danych jest priorytetem.
- Transformacja danych jest złożona i wymaga dokładnej kontroli przed załadowaniem.

Wybierz ELT, gdy:

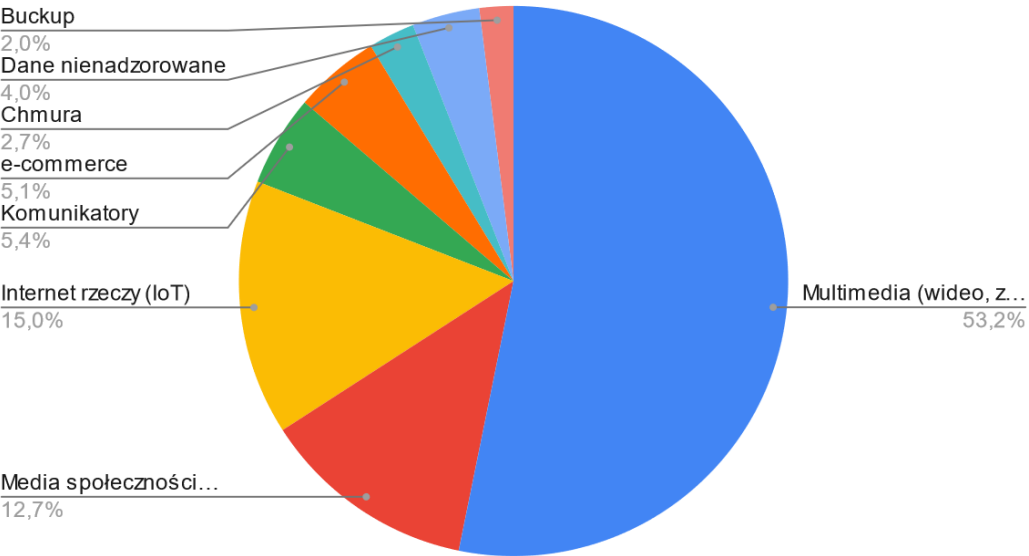
- Pracujesz z dużymi zestawami danych w środowisku Big Data lub chmurowym.
- Chcesz szybko ładować dane i wykonywać transformacje w systemie docelowym.
- Masz dostęp do dużej mocy obliczeniowej, która umożliwia efektywną transformację po załadowaniu danych.

Big Data

Wielkość danych w latach na Świecie



Typy danych



Na osobę	
2022*	12 TB
2023*	15 TB
2024*	18 TB
2025*	23 TB

Big Date

Big Date

to termin odnoszący się do ogromnych i złożonych zbiorów danych, które są zbyt duże, aby mogły być przetwarzane przez tradycyjne systemy baz danych.

Dane te charakteryzują się dużą objętością, różnorodnością i szybkością przetwarzania, co wymaga specjalistycznych technologii oraz narzędzi do analizy i zarządzania nimi.

Główne cechy (5 V)

Volume (Objętość):

- Ogromne ilości danych generowanych w krótkim czasie, często liczone w petabajtach lub eksabajtach.

Variety (Różnorodność):

- Dane dostępne są w różnych formatach, zarówno ustrukturyzowanych, półustrukturyzowanych, jak i nieustrukturyzowanych.

Velocity (Szybkość):

- Dane są generowane i przetwarzane w dużym tempie, często w czasie rzeczywistym, aby uzyskać natychmiastowe wyniki.

Veracity (Prawdziwość):

- Kluczowa jest jakość i wiarygodność danych, które muszą być dokładne, aby przyniosły wartościowe rezultaty.

Value (Wartość):

- Dane mają wartość tylko wtedy, gdy są odpowiednio przetworzone i zanalizowane, aby dostarczać użyteczne informacje wspierające decyzje.

Big Data

Zastosowanie

Marketing i personalizacja:

- Firmy mogą analizować dane dotyczące zachowań użytkowników online, aby dostosować reklamy i oferty do indywidualnych preferencji klientów.

Analiza predykcyjna:

- W branży finansowej i ubezpieczeniowej Big Data jest wykorzystywana do analizy ryzyka, przewidywania zmian rynkowych, a także w detekcji oszustw.

Zdrowie i medycyna:

- Analiza dużych zbiorów danych zdrowotnych pomaga w diagnozowaniu chorób, monitorowaniu pacjentów w czasie rzeczywistym oraz w rozwoju nowych terapii i leków.

IoT (Internet of Things):

- Urządzenia IoT generują ogromne ilości danych, które są analizowane w czasie rzeczywistym, aby zarządzać i optymalizować infrastrukturę, np. inteligentne miasta, systemy transportowe.

Media społecznościowe:

- Analiza danych z mediów społecznościowych pozwala na monitorowanie trendów, nastrojów konsumenckich, a także identyfikację wpływowych użytkowników.

Big Data

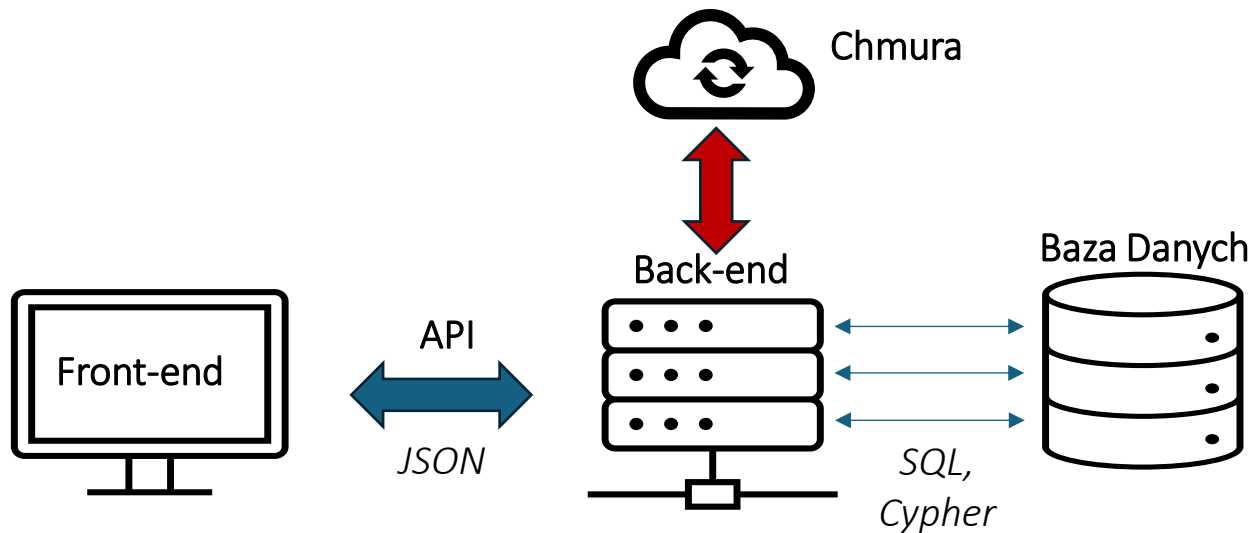
Big Data vs Hurtownia danych

Kryterium	Big Data	Hurtownia Danych
Rodzaj danych	Ustrukturyzowane, półustrukturyzowane, nieustrukturyzowane (np. dane tekstowe, multimedia).	Głównie ustrukturyzowane dane w tabelach relacyjnych.
Cel	Analiza dużych, złożonych zbiorów danych, analiza predykcyjna, wykrywanie wzorców i trendów.	Raportowanie i analiza historyczna, wspieranie decyzji biznesowych.
Skalowalność i rozmiar	Obsługuje ogromne ilości danych (petabajty i więcej), dane generowane w czasie rzeczywistym.	Obsługuje duże, ale ograniczone rozmiary danych, zazwyczaj z systemów informacyjnych organizacji.
Technologie	Hadoop, Apache Spark, NoSQL, chmura obliczeniowa.	RDBMS (Relacyjne Systemy Zarządzania Bazą Danych), ETL (Extract, Transform, Load).
Przetwarzanie danych	Przetwarzanie wsadowe i strumieniowe w czasie rzeczywistym.	Przetwarzanie wsadowe z okresowymi aktualizacjami (np. co dzień, tydzień).
Struktura danych	Złożone, dynamiczne, niejednolite dane.	Dane ustrukturyzowane.

Usługi chmurowe

Usługi Chmurowe

to szeroka gama rozwiązań IT, które są dostarczane przez Internet. Pozwalają na elastyczne zarządzanie zasobami IT, oferując skalowalność, elastyczność i możliwość płatności na zasadzie pay-as-you-go.



Usługi chmurowe

Zastosowania

Przechowywanie i Backup Danych:

- Usługi chmurowe oferują bezpieczne i skalowalne magazyny danych, które mogą być używane do przechowywania dużych ilości informacji oraz tworzenia kopii zapasowych.

Hosting Aplikacji Webowych:

- Platformy chmurowe umożliwiają hostowanie aplikacji internetowych i mobilnych, zapewniając wysoką dostępność i skalowalność.

Analiza Danych:

- Narzędzia do analizy danych w chmurze umożliwiają przetwarzanie i analizowanie dużych zbiorów danych w celu uzyskania wartościowych wniosków.

Uczenie Maszynowe i AI:

- Usługi chmurowe oferują zaawansowane narzędzia do trenowania i wdrażania modeli uczenia maszynowego i sztucznej inteligencji.

DevOps:

- Platformy chmurowe dostarczają narzędzia do ciągłej integracji i ciągłego dostarczania (CI/CD), umożliwiając automatyzację procesów programistycznych.

Usługi chmurowe

Chmura publiczna (Public Cloud)

Usługi są dostarczane przez zewnętrznego dostawcę i udostępniane publicznie.

Użytkownicy dzielą te same zasoby, co czyni ją bardziej ekonomiczną. Przykłady: Google Cloud, Amazon Web Services.



Chmura prywatna (Private Cloud)

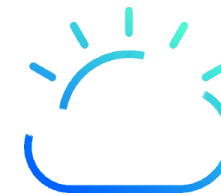
Infrastruktura chmury jest dedykowana jednemu użytkownikowi lub firmie, co zapewnia większą kontrolę nad bezpieczeństwem i prywatnością danych.

Może być hostowana na miejscu lub przez zewnętrznego dostawcę.



Usługi chmurowe

Główne platformy



Platforma	Opis	Główne Usługi
Amazon Web Services	AWS oferuje szeroki zakres usług chmurowych, od przetwarzania i magazynowania po sztuczną inteligencję i analitykę.	EC2, S3, RDS, Lambda, SageMaker
Microsoft Azure	Azure to kompleksowa platforma chmurowa Microsoftu, która oferuje usługi w zakresie przetwarzania, magazynowania, baz danych, IoT i wielu innych.	Virtual Machines, Azure SQL Database, Azure Functions, Azure DevOps
Google Cloud Platform	GCP dostarcza usługi chmurowe Google, w tym zaawansowane usługi analizy danych, uczenia maszynowego, przetwarzania i przechowywania danych.	Compute Engine, Cloud Storage, BigQuery, Cloud Functions
IBM Cloud	IBM Cloud oferuje zarówno publiczne, prywatne, jak i hybrydowe rozwiązania chmurowe, z usługami obejmującymi AI, blockchain, IoT i wiele innych.	IBM Cloud Kubernetes Service, IBM Watson, IBM Blockchain Platform
Oracle Cloud	Oracle Cloud oferuje szeroką gamę usług chmurowych, szczególnie skoncentrowanych na rozwiązaniach baz danych i aplikacjach biznesowych.	Oracle Autonomous Database, Oracle Cloud Infrastructure, Oracle Cloud Applications

Humans are HOOKED

Machines are LEARNING

