



HURTOWNIE DANYCH

Krzysztof Goczyła

Wydział Elektroniki, Telekomunikacji i Informatyki

Politechnika Gdańska

kris@eti.pg.gda.pl





Część I.

Tworzenie hurtowni danych

1. Co to jest hurtownia danych?
2. Model danych w hurtowni danych
3. Przykłady hurtowni danych i analiz
4. Architektura logiczna hurtowni danych
5. Architektura fizyczna hurtowni danych
6. Obszary zastosowań
7. Planowanie hurtowni danych

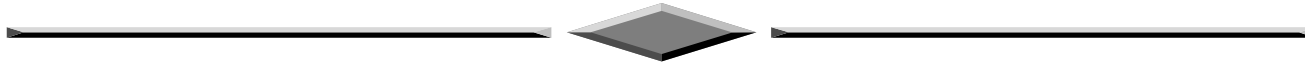
Literatura:

- V. Poe, P.Klauer, S.Brebst. „Tworzenie hurtowni danych”, WNT 2000
R. Kimball. „Data Warehouse Toolkit”. J. Wiley&Sons, 1996.
W.H.Inmon. „Building the Data Warehouse”. J. Wiley&Sons, 2002.



Co to jest hurtownia danych?

- **Scentralizowana nietransakcyjna baza danych** przeznaczona do przechowywania informacji w długim horyzoncie czasowym **globalnie** w skali instytucji, w **wielowymiarowych układach analitycznych** i ukierunkowana na wyszukiwanie i analizowanie informacji bezpośrednio przez końcowych użytkowników.



- Tematyczny, zintegrowany, zależny od czasu, trwały zbiór danych, ukierunkowany na **wspomaganie procesów podejmowania decyzji**.

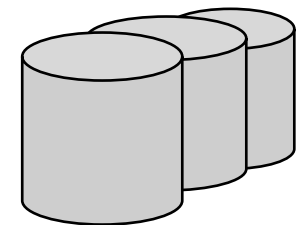


Cechy hurtowni danych

Baza danych

Hurtownia danych to bardzo duża baza danych (setki GB), przechowująca dane z długiego horyzontu czasowego.

Taka baza danych optymalizowana jest pod kątem przetwarzania analitycznego, a nie transakcyjnego.





Cechy hurtowni danych

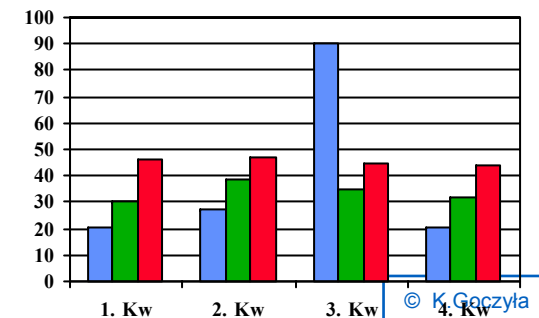
Przetwarzanie nietransakcyjne

Operacje dokonywane na hurtowni danych:

- nie zmieniają zawartości bazy danych,
- wydobywają informacje w różnych przekrojach i agregacjach.

Przetwarzanie typu **OLAP** (*On-Line Analytical Processing*):

Przetwarzanie danych, którego celem są analizy trendów, analizy przekrojowe i inne analizy o charakterze strategicznym





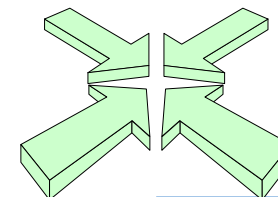
Cechy hurtowni danych

Scentralizowanie

Dane pochodzące z wielu różnych systemów baz danych zbierane są do jednego miejsca (**scentralizowanej bazy danych**), gdzie rezyduje hurtownia danych.

W tym miejscu realizowane są:

- obróbka danych
- analizy
- prezentacja danych i wyników





Cechy hurtowni danych

Globalność

- Hurtownia danych obejmuje *całe* przedsiębiorstwo (organizację, instytucję, ...).
- Zawiera wszystkie, kompletne dane dotyczące określonej dziedziny działalności przedsiębiorstwa (w przeciwnym razie wyniki analiz OLAP mogą nie być miarodajne).



Gdy dane w hurtowni obejmują tylko pewien wycinek danych globalnych:



minihurtownia

(podhurtownia, zbiorcza baza danych, *data mart*)



Cechy hurtowni danych

Wspomaganie podejmowania decyzji

Hurtownie danych - dobra podstawa do tworzenia systemów wspomagania decyzyjnego (DSS, *Decision Support System*):

- silne narzędzia analityczne
- wydajne przetwarzanie ogromnych ilości danych

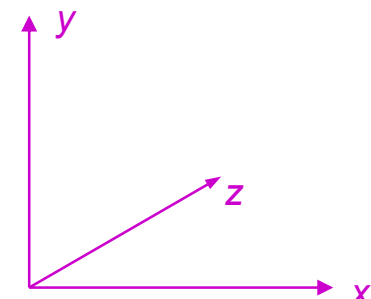




Cechy hurtowni danych

Wielowymiarowy model danych

- Zasadnicze dane hurtowni przechowywane są w postaci **faktów** o charakterze numerycznym, mogących być przedmiotem analiz ilościowych.
- **Wymiary** o charakterze nienumerycznym (opisowym) służą do agregowania faktów względem różnych kryteriów (warunków określonych na wymiarach).





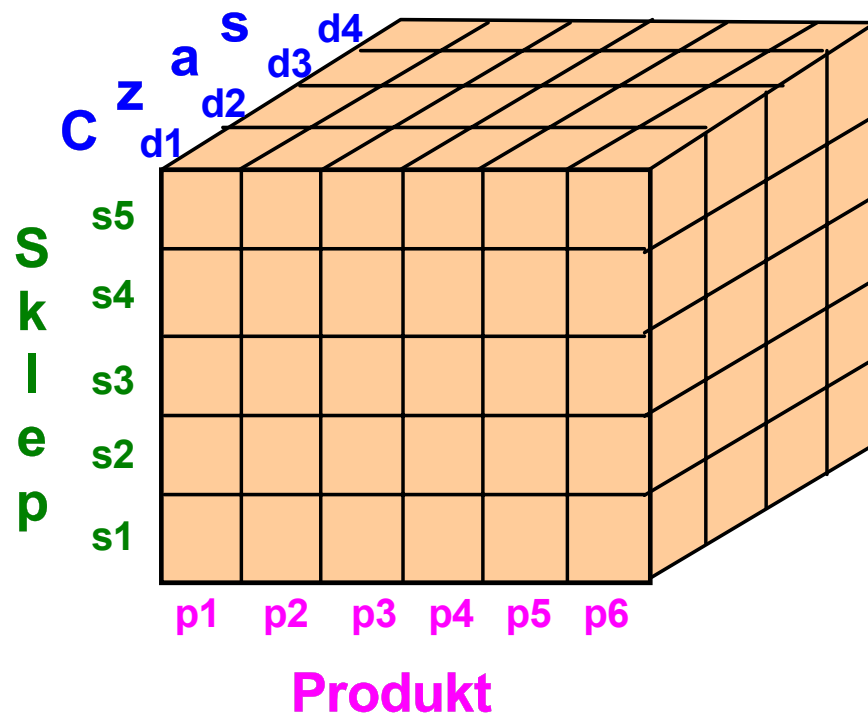
Cechy hurtowni danych - podsumowanie

- Bardzo duża baza danych
- Ładowana z zewnętrznych źródeł danych
- Przeznaczona tylko do odczytu
- Zorganizowana pod kątem analiz przekrojowych



Wielowymiarowy model danych

Sieć sprzedaży



Produkt

Sklep

Czas

wymiary

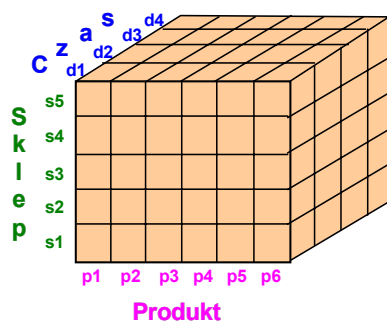
(p_i, s_i, d_i)

fakt sprzedaży
produktu p_i
w sklepie s_i
dnia d_i



Przykłady analiz przekrojowych

Wycinanie



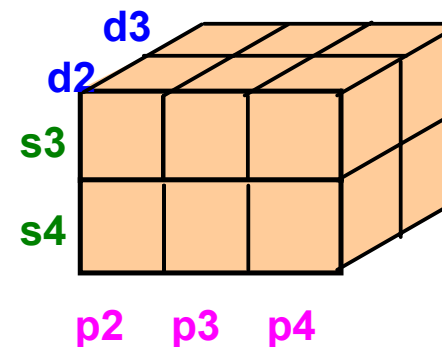
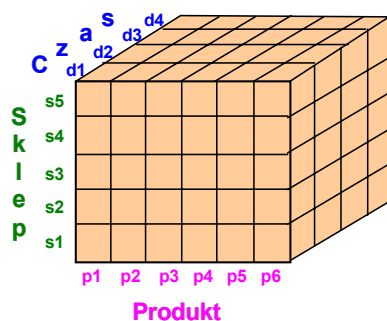
Efekt:

(p_i, s_3, d_j) - wszystkie fakty sprzedaży w sklepie s_3



Przykłady analiz przekrojowych

Wycinanie



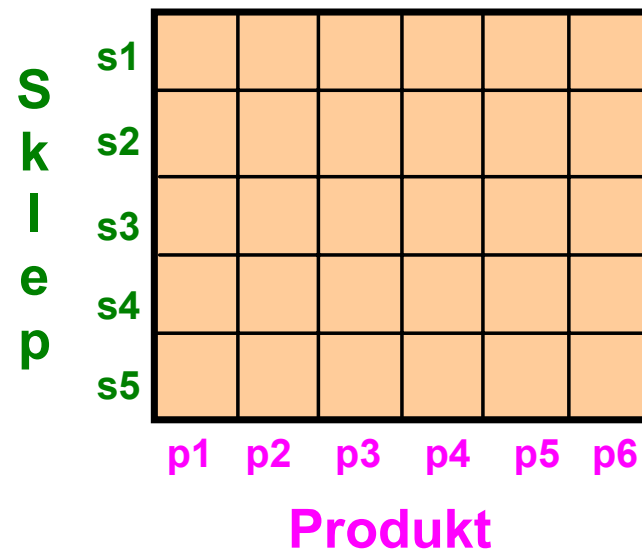
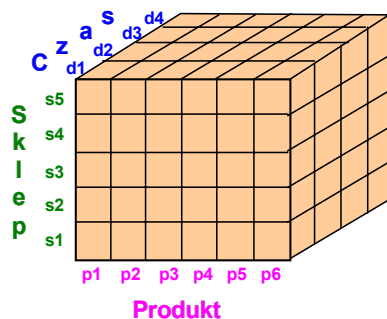
Efekt:

(p_i, s_i, d_i) - fakty sprzedaży w sklepie s_3 i s_4
produktów p_2, p_3, p_4 w dniach d_2, d_3



Przykłady analiz przekrojowych

Rzutowanie

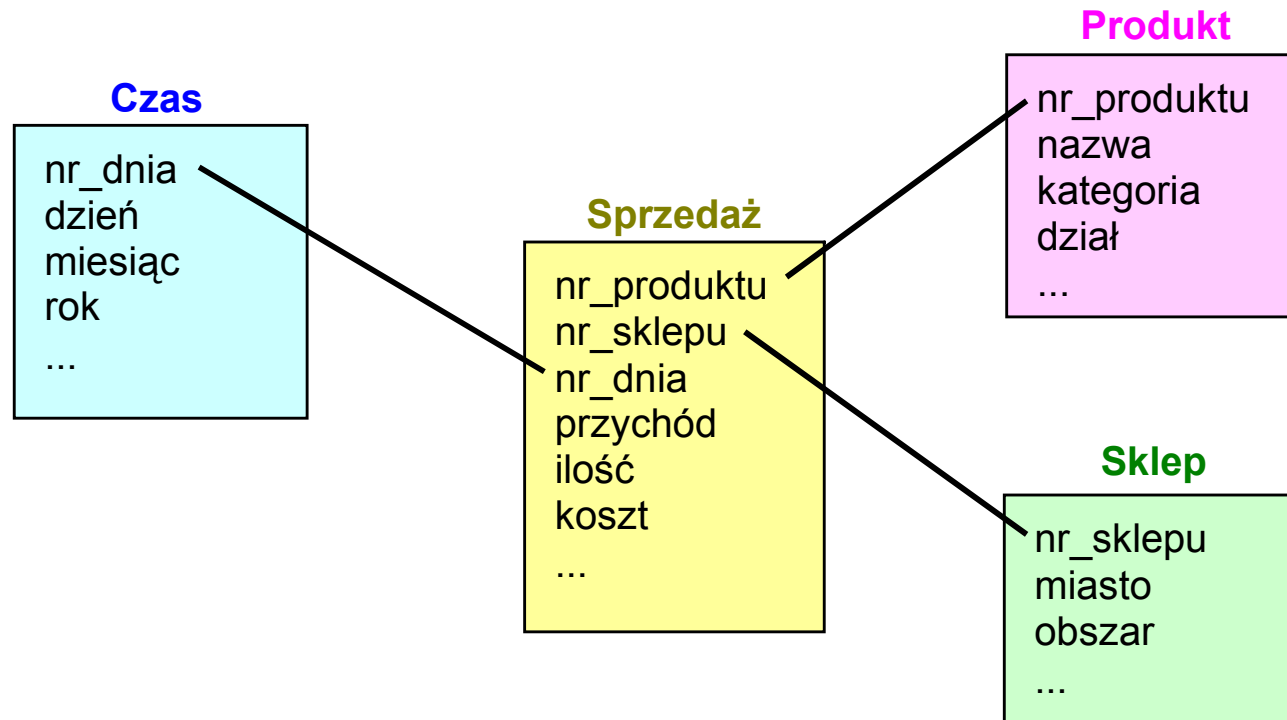


Efekt:

(p_i, s_i) - agregacja sprzedaży poszczególnych produktów w poszczególnych sklepach w całym okresie



Schemat gwiazdy



Sprzedaż - tablica faktów, z atrybutami *liczbowymi*

Czas, **Produkt**, **Sklep** - tablice wymiarów, z atrybutami *opisowymi*

nr_xxx... to atrybuty *kluczowe*



Przykłady zapytań analitycznych

Podaj wielkość sprzedaży (kwotowo i ilościowo) w roku 1998 wszystkich produktów z poszczególnych działów, we wszystkich sklepach.

Wynik:

Dział	Przychód	Ilość
chemiczne	1345,90	3567
elektryczne	9878,00	456
papiernicze	6784,35	1765
spożywcze	12456,20	10345

Zapytanie SQL:

```
SELECT p.dział, SUM(s.przychód), SUM(s.ilosc)
FROM Sprzedaż s, Produkt p, Czas c
WHERE s.nr_produktu = p.nr_produktu AND
      s.nr_dnia = c.nr_dnia AND
      c.rok = 1998
GROUP BY p.dział
ORDER BY p.dział
```



Przykłady zapytań analitycznych

Podaj wielkość sprzedaży (kwotowo i ilościowo) w roku 1998 wszystkich produktów z poszczególnych działów i kategorii, we wszystkich sklepach.

Wynik:

Dział	Kategoria	Przychód	Ilość
chemiczne	farby	740,60	2557
chemiczne	proszki	605,30	1010
elektryczne	wtyczki	1550,50	123
elektryczne	żarówki	8327,50	333
papiernicze	piśmienne	684,00	1000
papiernicze	zeszyty	6100,35	765
spożywcze	mleczne	9500,10	7500
spożywcze	pieczywo	956,10	2500
spożywcze	wędliny	2000,00	345

Dodano jeden atrybut wymiaru **Produkt: kategoria**, uszczegóławiając obraz danych.



rozwijanie danych (*drilling down*)



Przykłady zapytań analitycznych

Zapytanie SQL:

```
SELECT p.dział, p.kategoria, SUM(s.przychód), SUM(s.ilosc)
FROM Sprzedaż s, Produkt p, Czas c
WHERE s.nr_produktu = p.nr_produktu AND
      s.nr_dnia = c.nr_dnia AND
      c.rok = 1998
GROUP BY p.dział, p.kategoria
ORDER BY p.dział, p.kategoria
```

Poprzednio:

```
SELECT p.dział, SUM(s.przychód), SUM(s.ilosc)
FROM Sprzedaż s, Produkt p, Czas c
WHERE s.nr_produktu = p.nr_produktu AND
      s.nr_dnia = c.nr_dnia AND
      c.rok = 1998
GROUP BY p.dział
ORDER BY p.dział
```



Przykłady zapytań analitycznych

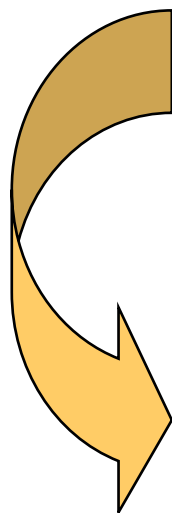
Dział	Przychód	Ilość
chemiczne	1345,90	3567
elektryczne	9878,00	456
papiernicze	6784,35	1765
spożywcze	12456,20	10345

zwijanie
(drilling up)



Dział	Kategoria	Przychód	Ilość
chemiczne	farby	740,60	2557
chemiczne	proszki	605,30	1010
elektryczne	wtyczki	1550,50	123
elektryczne	żarówki	8327,50	333
papiernicze	piśmienne	684,00	1000
papiernicze	zeszyty	6100,35	765
spożywcze	mleczne	9500,10	7500
spożywcze	pieczywo	956,10	2500
spożywcze	wędliny	2000,00	345

rozwijanie
(drilling down)





Przykłady zapytań analitycznych

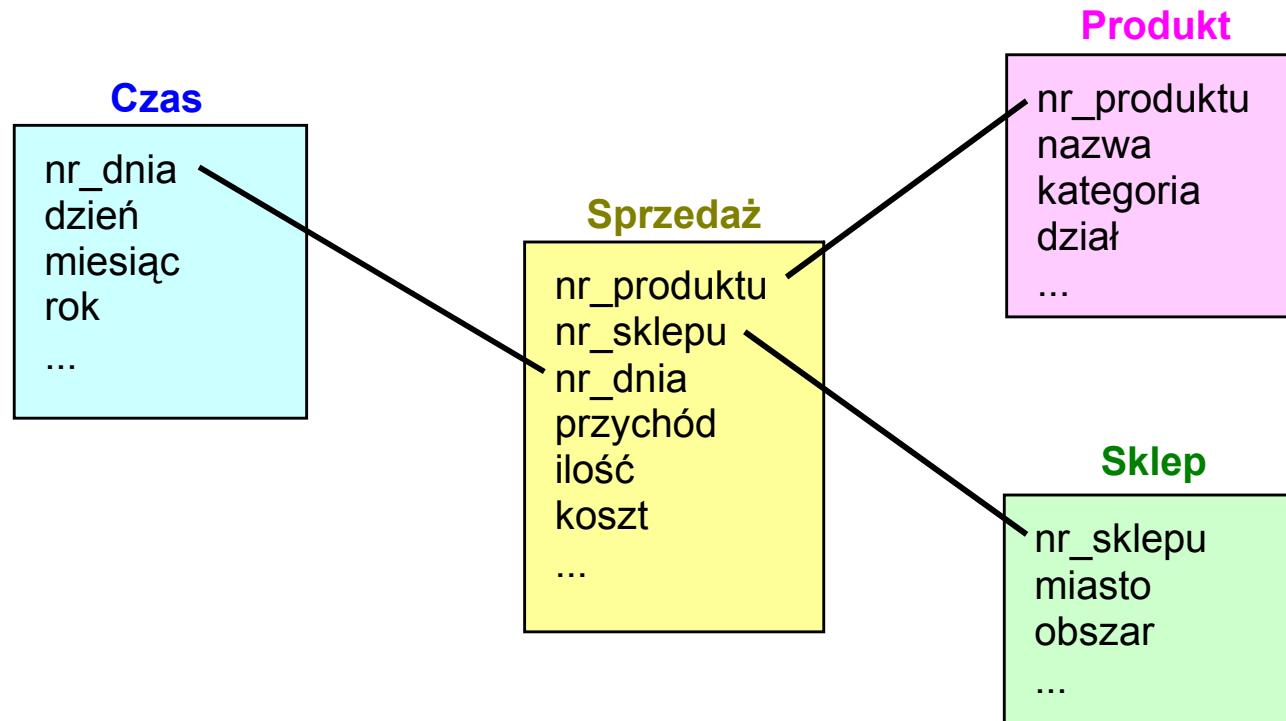
Podaj zestawienie sprzedaży w 1998 roku według działów produktów, z dokładnością do miesiąca.

Dział	Miesiąc	Przychód	Ilość
chemiczne	styczeń	34,10	12
chemiczne	luty	120,00	40
...
chemiczne	grudzień	20,50	10
elektryczne	styczeń	321,90	87
elektryczne	luty	421,00	101
...
papiernicze	styczeń	145,20	97
...

Dział i Miesiąc to atrybuty *różnych* wymiarów



Schemat gwiazdy



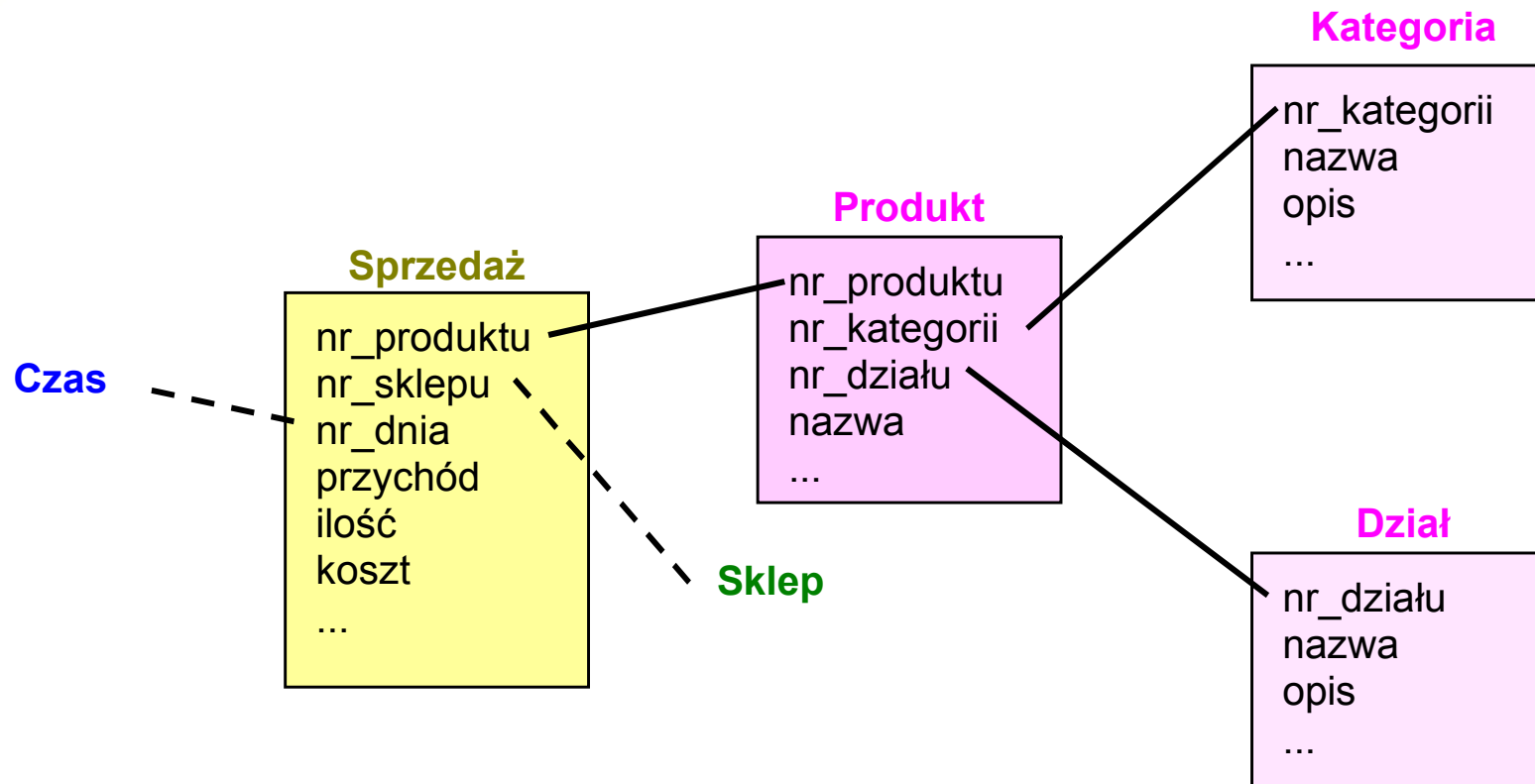
Niektóre atrybuty wymiarów mogą się wielokrotnie powtarzać
(np. **miesiąc**, **rok**, **kategoria**, **dział**, **obszar**)



redundancja danych



Schemat płatka śniegu



- Pozwala usunąć redundancję i zredukować wielkość bazy danych
- Zmniejsza efektywność realizacji zapytań analitycznych



Modele pamięci w hurtowniach danych

- ROLAP (Relational OLAP)

Wszystkie dane i agregaty przechowywane są w tablicach relacyjnej bazy danych (często – w postaci źródłowej).

+ **nie potrzeba dodatkowej pamięci**

- **słaba efektywność**

- MOLAP (Multidimensional OLAP)

Wszystkie dane źródłowe ładowane są do specjalnych struktur wielowymiarowych, zoptymalizowanych pod kątem przetwarzania analitycznego. Wstępnie obliczane są agregaty.

+ **najlepsza efektywność**

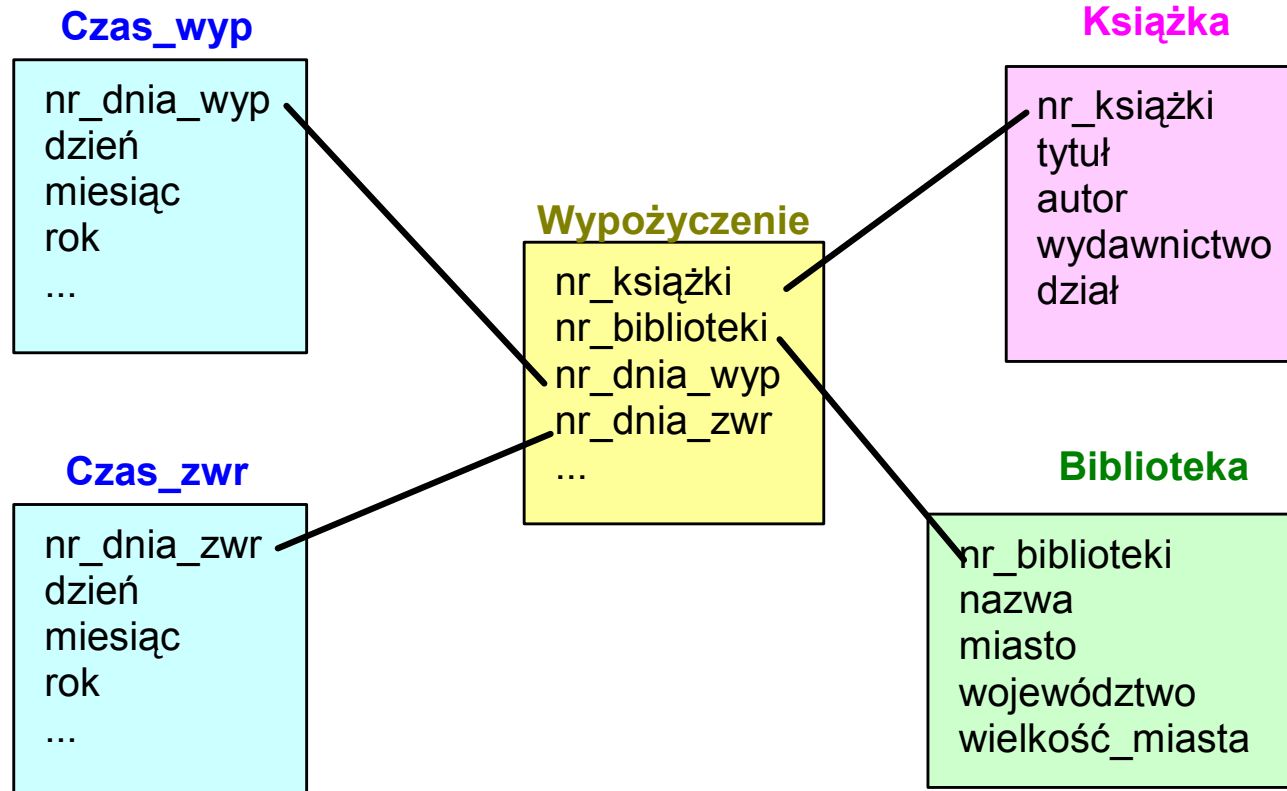
- **wymaga dużo dodatkowej pamięci**

- HOLAP (Hybrid OLAP)

Rozwiązanie pośrednie: dane źródłowe przechowywane są w tablicach, natomiast agregaty są wstępnie obliczane i przechowywane w specjalizowanych strukturach wielowymiarowych.



Model hurtowni dla sieci bibliotek - fakty i miary



- Fakt **Wypożyczenie** może nie mieć żadnych atrybutów liczbowych; wszelkie analizy będą bazować na *liczbie* faktów. Możemy sobie wyobrazić, że miarą jest liczba 1.



Model hurtowni dla sieci bibliotek – wymiary

- Wymiary **Czas_wyp** i **Czas_zwr** mogą być implementowane za pomocą jednej tablicy **Czas** o kluczu **nr_dnia**.
- Z jednej tablicy wymiarów można utworzyć wiele wymiarów:
 - **Wydawnictwo**: ▪ **Książka.wydawnictwo**
 - **DziałLiteracki**: ▪ **Książka.dział**
 - **Wielkość**: ▪ **Biblioteka.wielkość_miasta**
 - **Położenie**: ▪ **Biblioteka.województwo**
 - ▪ **Biblioteka.miasto**
 - ▪ ▪ **Biblioteka.nazwa**

Położenie jest wymiarem hierarchicznym;
Wydawnictwo, **DziałLiteracki** i **Wielkość** to wymiary kategorijskie.

Zazwyczaj wymiar oznaczający czas jest wymiarem hierarchicznym; np.

- **Czas_wyp**: ▪ **Czas.miesiąc**
 - ▪ **Czas.rok**



Przykłady zapytań analitycznych

Podaj liczby wypożyczeń w roku 1998 według miast, uszeregowane w kolejności malejącej.

Wynik:

Miasto	Ile
Gdańsk	13667
Warszawa	10234
Poznań	9765
Wrocław	9345
Kraków	8231
...	...

Zapytanie SQL:

```
SELECT b.miasto, COUNT(*) AS Ile
FROM Wypożyczenie w, Biblioteka b, Czas c
WHERE w.nr_biblioteki = b.nr_biblioteki AND
      w.nr_dnia_wyp = c.nr_dnia AND
      c.rok = 1998
GROUP BY b.miasto
ORDER BY Ile
```



Przykłady zapytań analitycznych

Rozwiń poprzednie zapytanie, uszczegóławiając je o działy książek.

Wynik:

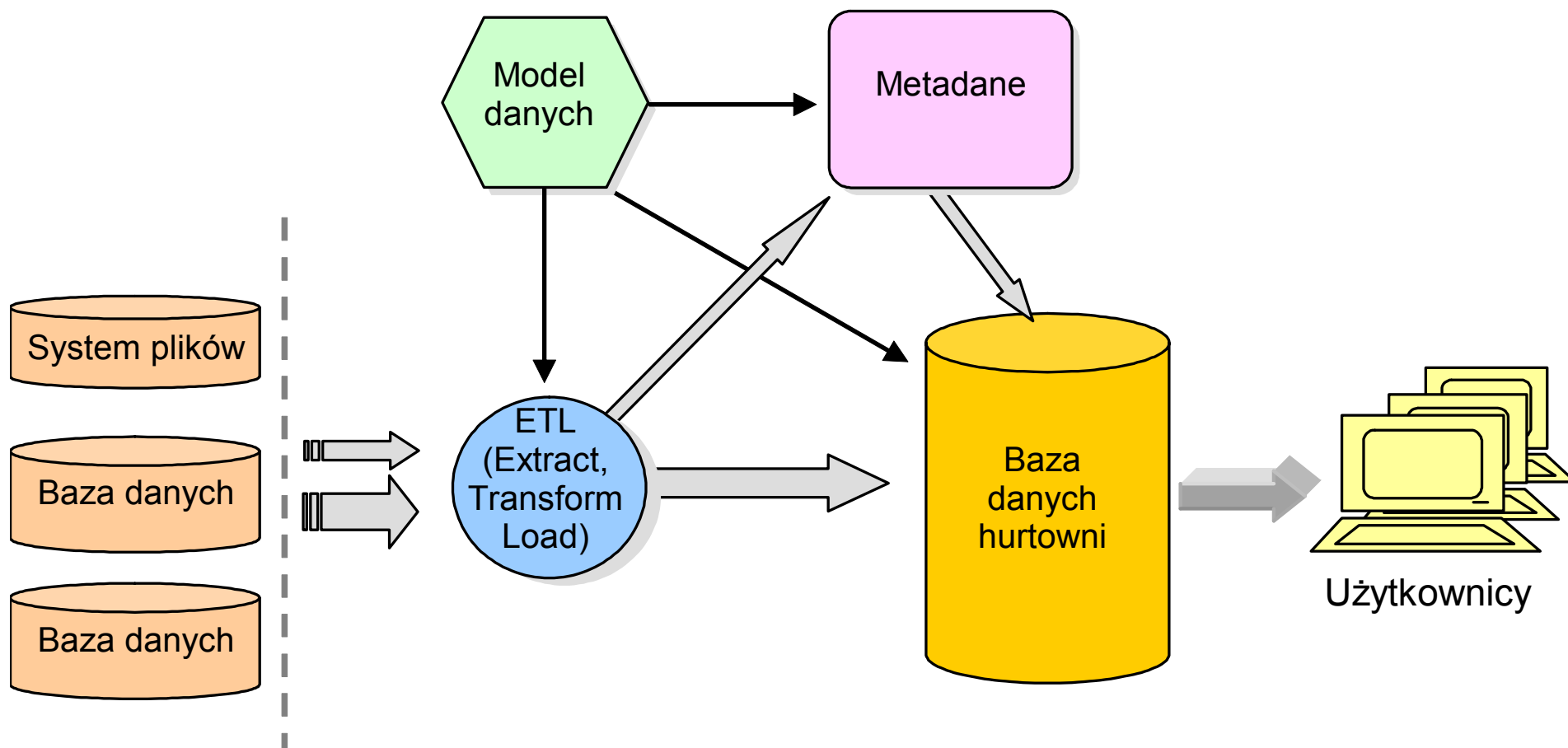
Miasto	Dział	Ile
Gdańsk	beletr.	678
Gdańsk	naukowe	100
Gdańsk	słowniki	322
...
Kraków	beletr.	456
Kraków	naukowe	99
...
Poznań	beletr.	377
...
Warszawa	beletr.	477
...

Zapytanie SQL:

```
SELECT b.miasto, k.dział, COUNT(*) AS Ile
FROM Wypożyczenie w, Biblioteka b, Czas c, Książka k
WHERE w.nr_biblioteki = b.nr_biblioteki AND
      w.nr_dnia_wyp = c.nr_dnia AND
      w.nr_książki = k.nr_książki AND
      c.rok = 1998
GROUP BY b.miasto, k.dział
ORDER BY b.miasto, k.dział
```

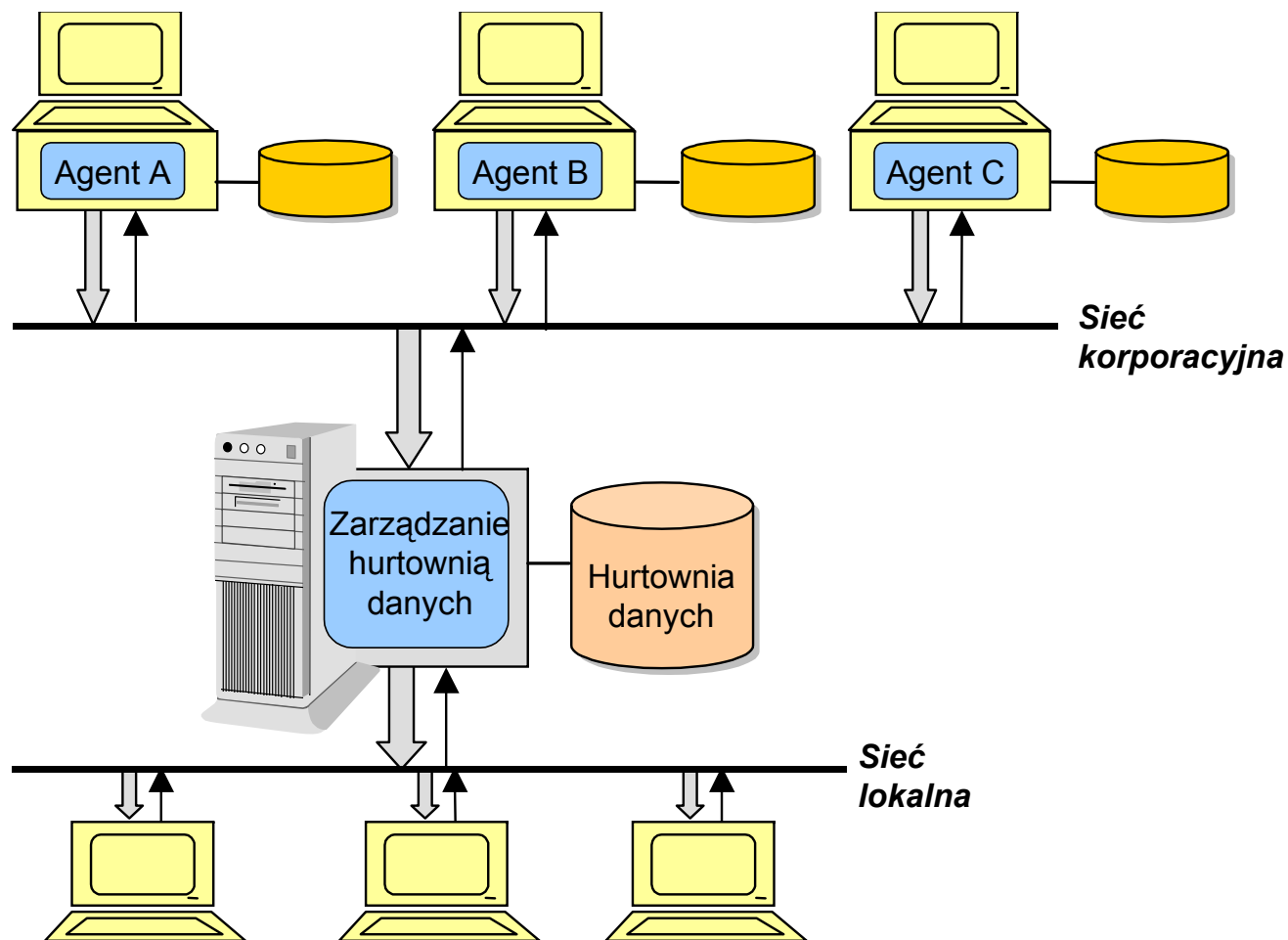


Architektura logiczna hurtowni danych





Architektura fizyczna hurtowni danych





Obszary zastosowań

Dziedzina	Fakty
<ul style="list-style-type: none">• Sieć sprzedaży detalicznej	<ul style="list-style-type: none">• sprzedaż
<ul style="list-style-type: none">• Sieć hurtowni	<ul style="list-style-type: none">• dostawa• wysyłka
<ul style="list-style-type: none">• Operator telekomunikacyjny	<ul style="list-style-type: none">• połączenie
<ul style="list-style-type: none">• Bank	<ul style="list-style-type: none">• operacja finansowa
<ul style="list-style-type: none">• Instytucja ubezpieczeniowa	<ul style="list-style-type: none">• umowa• szkoda
<ul style="list-style-type: none">• Linie lotnicze	<ul style="list-style-type: none">• przelot
<ul style="list-style-type: none">• Sieć meteo	<ul style="list-style-type: none">• pomiar



Producenci i produkty

- Serwery OLAP
 - Hurtownie danych i narzędzia analityczne
-

Oracle Corporation

serwery Oracle 10g
narzędzia OLAP

IBM Corporation

serwery DB2
hurtownia Visual Warehouse

Informix Software Inc.

serwery OnLine Dynamic Server
hurtownia Metacube

Microsoft Corporation

serwer MS SQL 2005

Sybase Inc.

serwery Sybase IQ

Arbor Software Corporation

serwery Essbase Hyperion

Red Brick Systems Inc.

hurtownia Red Brick Warehouse

SAS Institute

pakiet SAS System

Cognos Inc.

narzędzia CognosSuite
(*business intelligence*)



Planowanie hurtowni danych

1. Jakie informacje są potrzebne do podejmowania decyzji na poziomie strategicznym?
2. Czy odpowiednie dane są aktualnie gromadzone w miejscach działalności? Jeśli nie, to jakie nakłady są potrzebne, by je gromadzić?
3. Określ, jakiego rodzaju analizy danych będą potrzebne do podejmowania decyzji na poziomie strategicznym.
4. Zaprojektuj hurtownię danych (fakty, wymiary). Może potrzebnych jest kilka kostek?
5. Wybierz serwer OLAP i narzędzia *business intelligence*. Określ niezbędną konfigurację sprzętową i programową (**koszty!**).
6. Zidentyfikuj formaty danych, jakie są gromadzone w poszczególnych miejscach działalności.
7. Opracuj procedury przekazywania danych źródłowych do hurtowni danych. Określ sposób traktowania wartości brakujących i odstających.
8. Zaimplementuj hurtownię danych dla określonego wycinka działalności (np. dla jednego rodzaju usług, dla jednego obszaru geograficznego itp.).
9. Oceń efekty poprzedniego kroku. Podejmij decyzję o wdrożeniu globalnym.
10. Starannie monitoruj funkcjonowanie hurtowni.



Planowanie hurtowni danych – cd.

- **Jakie informacje są potrzebne do podejmowania decyzji na poziomie strategicznym?**
 - Jaki jest cel zbierania informacji?
 - Kto jest odbiorcą raportów?
 - Zakres obowiązków
 - Dane analizowane – źródła
 - Dane sporządzane – odbiorcy
 - Lista życzeń
 - Jakie raporty są wymagane przez każdego z odbiorców?
- **Czy odpowiednie dane są aktualnie gromadzone w miejscach działalności? Jeśli nie, to jakie nakłady są potrzebne, by je gromadzić?**
 - Jakie systemy działają w przedsiębiorstwie?
 - Jakie dane gromadzi każdy z systemów?
 - Jakie dane są przekazywane obecnie między systemami?
 - Jakie są potrzeby? Co trzeba zrobić, by je spełnić?



Planowanie hurtowni danych – cd.

- **Określ, jakiego rodzaju analizy danych będą potrzebne do podejmowania decyzji na poziomie strategicznym**
 - Jak są cele strategiczne przedsiębiorstwa?
 - Co jest niezbędne, aby je zrealizować?
 - Co może w tym przeszkodzić?
 - Jak można wykorzystać dane zbierane w systemach do osiągnięcia tych celów?
- **Zaprojektuj hurtownię danych (fakty, wymiary). Może potrzebnych jest kilka kostek?**
 - Zidentyfikuj fakty.
 - Dla każdego faktu określ wymiary.
 - Każdy fakt i każdy wymiar skojarz ze źródłem danych.
 - Zdefiniuj kostki.