Hurtownie danych a transakcyjne bazy danych

Materiały źródłowe do wykładu:

[1] Jerzy Surma, *Business Intelligence. Systemy wspomagania decyzji*, Wydawnictwo Naukowe PWN, Warszawa 2009

[2] Arkadiusz Januszewski, *Funkcjonalność informatycznych systemów zarządzania, Tom 2: Systemy Business Intelligence*, Wydawnictwo Naukowe PWN, Warszawa 2008

[3] Materiały szkoleniowe Oracle, w większości dostępne w ramach licencji Oracle Academy:

- Oracle Database 11g: Implement and Administer a Data Warehouse
- Oracle10g: Data Warehousing Fundamentals
- Oracle Warehouse Builder 10g: Implementation Part I and Part II
- Data Modeling and Relational Database Design

Hurtownie danych a transakcyjne bazy danych

Główne kategorie systemów baz danych:

- Transakcyjne
  OLTP (*OnLine Transaction Processing*) – obecnie
  głównie relacyjne bądź relacyjno-obiektowe bazy danych,
  zoptymalizowane na sprawną obsługę wielu transakcji
  często wykonywanych przez wielu użytkowników

- Analityczne
  OLAP (*OnLine Analitical Processing*) – hurtownie
  danych (DW – *Data Warehouse*), współpracujące
  z narzędziami eksploracji danych (DM – *Data Mining*)
  i systemami wspomagania decyzji (DSS – *Decision Support System*), które obecnie rozbudowano do zaawansowanych systemów analityki biznesowej (BI – *Business Intelligence*)

Definicja hurtowni danych

Hurtownia danych to [1]:

- uporządkowany tematycznie (*subject oriented*),
- zintegrowany (*integrated*),
- zawierający wymiar czasowy (*time variant*),
- nieulotny (*non-volatile*)

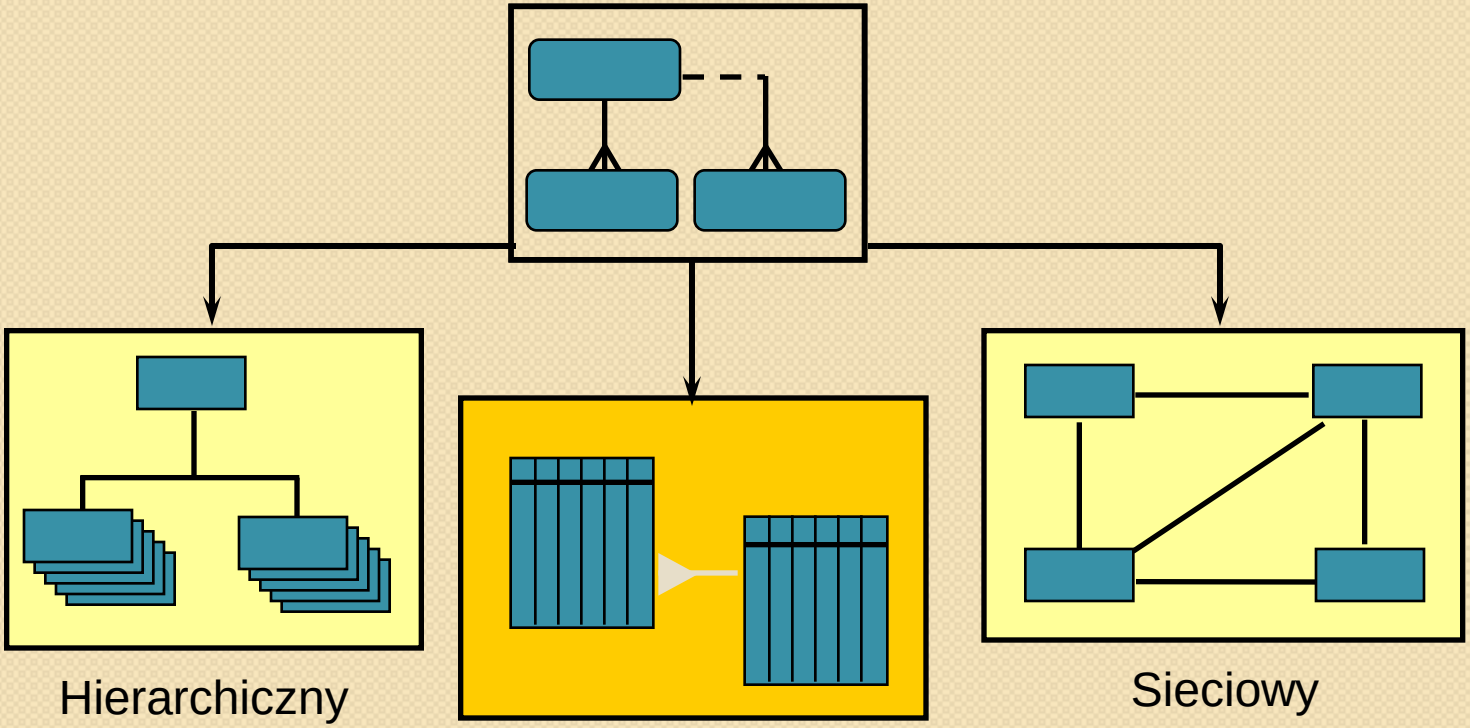
zbiór danych (baza danych) wspomagających podejmowanie decyzji.

Porównanie OLTP i hurtowni danych

Parametr	OLTP	Hurtownia danych
Czas odpowiedzi	Sekundy i krótszy	Sekundy i godziny
Organizacja danych	Zwykle podporządkowana konkretnej aplikacji	Zdeterminowana tematem i czasem
Aktywność	Procesy	Analizy
Połączenia	Wiele	Niewiele
Natura danych	Transakcyjne, aktualne, zwykle przechowywane 30-60 dni	Historyczne, migawki czasowe, dane poddane obróbce, agregaty
Rozmiar danych	Od małego do dużego	Duży do bardzo dużego
Źródła danych	Operacyjne, wewnętrzne	Operacyjne, wewnętrzne, zewnętrzne
Normalizacja	Znormalizowane	Zdenormalizowane

Modelowanie systemów OLTP

Model związków encji
ER Model – *Entity Relationship Model* [3]



Hierarchiczny

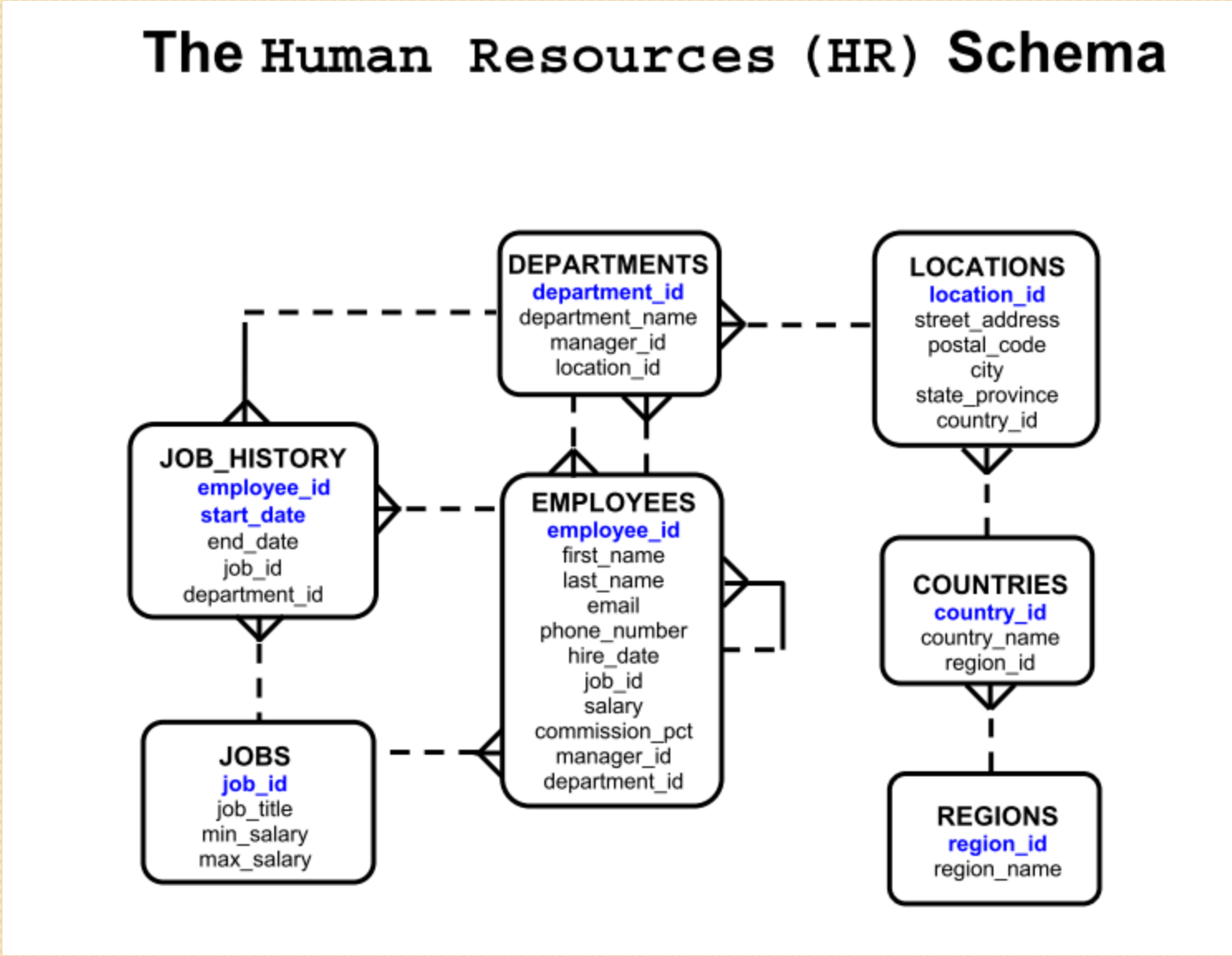
Relacyjny
lub
relacyjno-obiektowy

Sieciowy

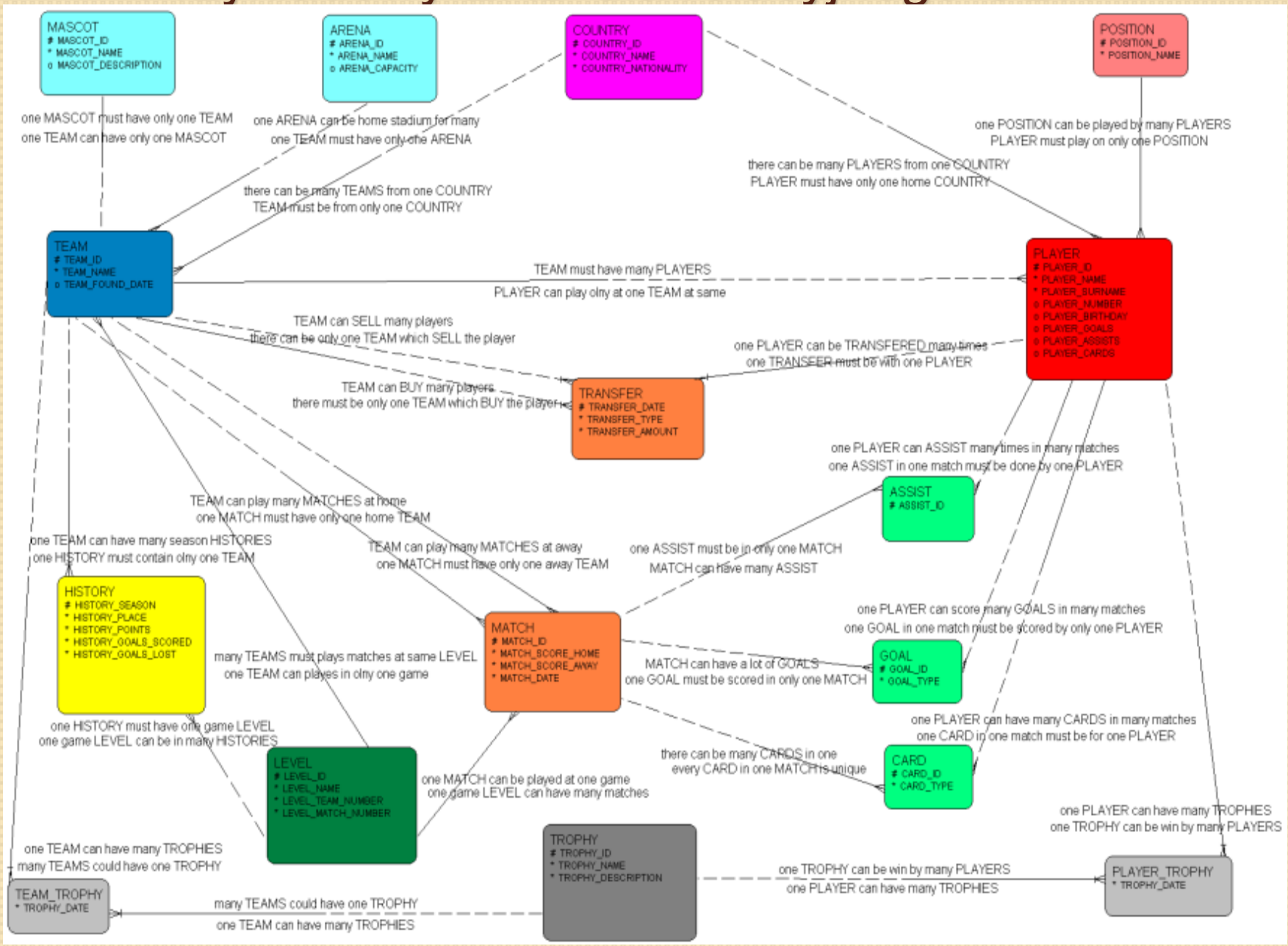
Normalizacja - podstawa dobrego relacyjnego OLTP

- 1NF (1st Normal Form) – pierwsza postać normalna relacji R bazy danych wymaga spełnienia warunku atomowości w relacji R dla każdego atrybutu w każdej krotce
- 2NF (2nd Normal Form) – druga postać normalna relacji R bazy danych wymaga 1NF i aby nie było częściowych zależności funkcyjnych atrybutów niekluczowych relacji R od klucza tej relacji.
- 3NF (3rd Normal Form) – trzecia postać normalna relacji R bazy danych wymaga 2NF i aby nie było przechodnich zależności funkcyjnych atrybutów niekluczowych relacji R od klucza tej relacji
- 4NF (4th Normal Form) – czwarta postać normalna relacji R bazy danych wymaga 3NF i dozwala co najwyżej jedną wielowartościową zależność funkcyjną w tej relacji.

Przykładowy ERD dla relacyjnego OLTP [3]



Przykładowy ERD dla relacyjnego OLTP



Klasyfikacja systemów OLAP-DW

- Wielowymiarowe

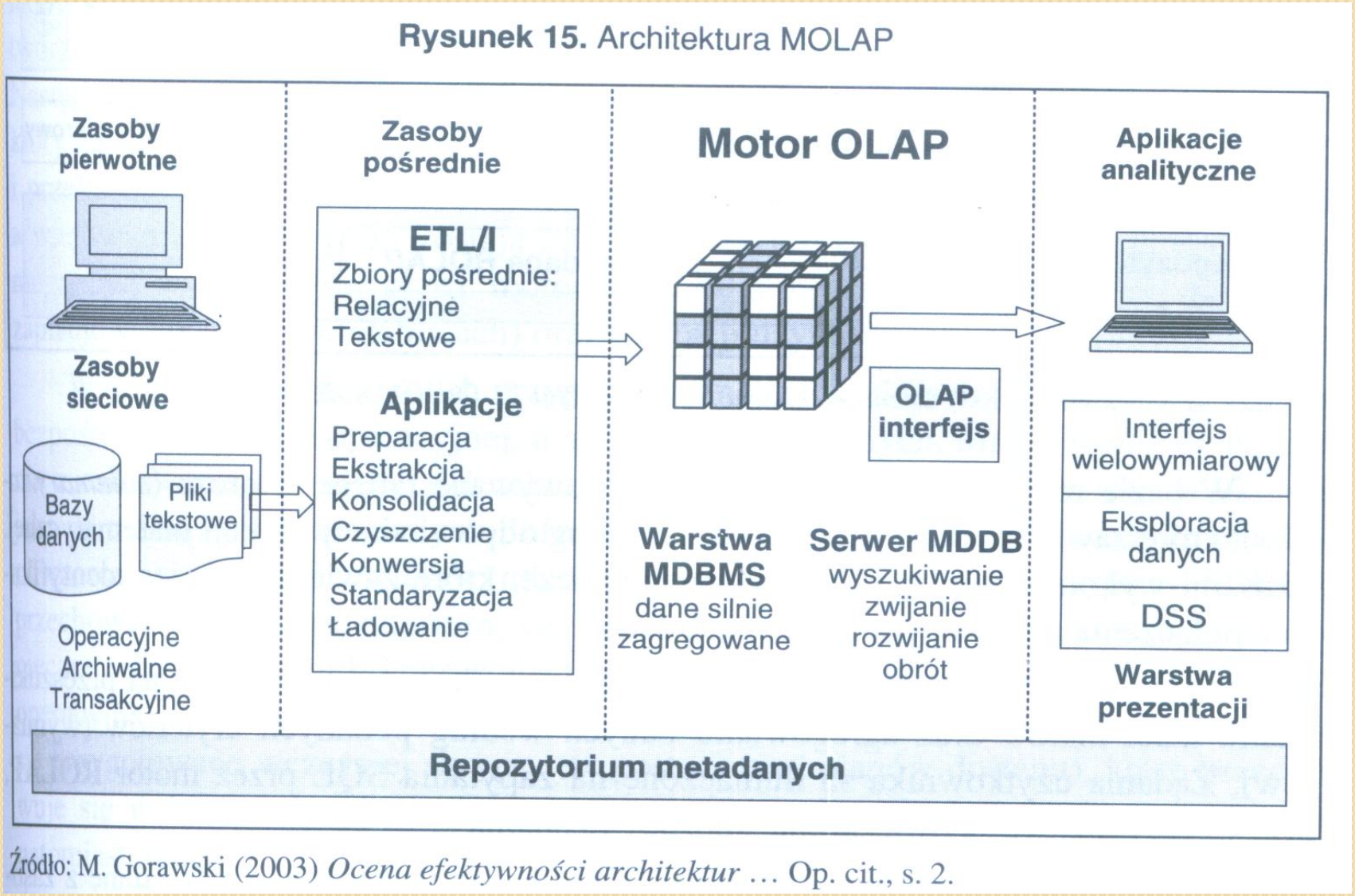
  MOLAP (*Multidimensional OLAP*) – tradycyjne rozwiązanie sprowadzające się do składowania danych w postaci zagregowanej w tzw. wielowymiarowych kostkach (*Multidimensional Cubes*)

- Relacyjne

  ROLAP (*Relational OLAP*) – oparte na relacyjnych bazach danych, w których oprócz danych źródłowych składowane są tzw. tabele faktów (*Fact Tables*), pozwalające konstruować wielowymiarowe kostki w zależności od potrzeb analitycznych

- Hybrydowe – łączą w sobie cechy rozwiązań MOLAP i ROLAP

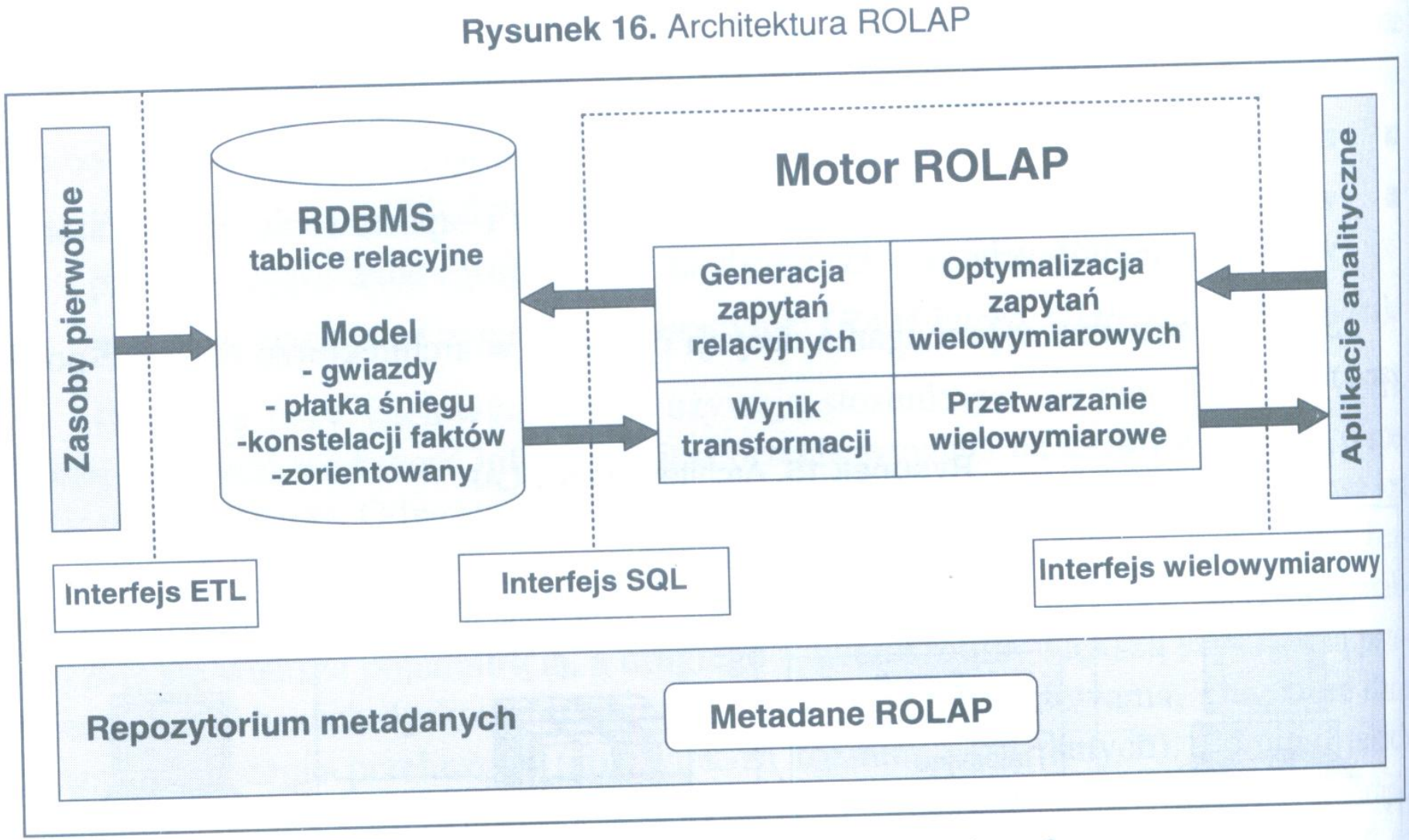
Architektura MOLAP [2]



Źródło: M. Gorawski (2003) *Ocena efektywności architektur …* Op. cit., s. 2.

ETL/I (Extraction, Transformation, Loading Interface)
– Interfejs ekstrakcji, transformacji i ładowania danych do hurtowni

Architektura ROLAP [2]



Rysunek 16. Architektura ROLAP

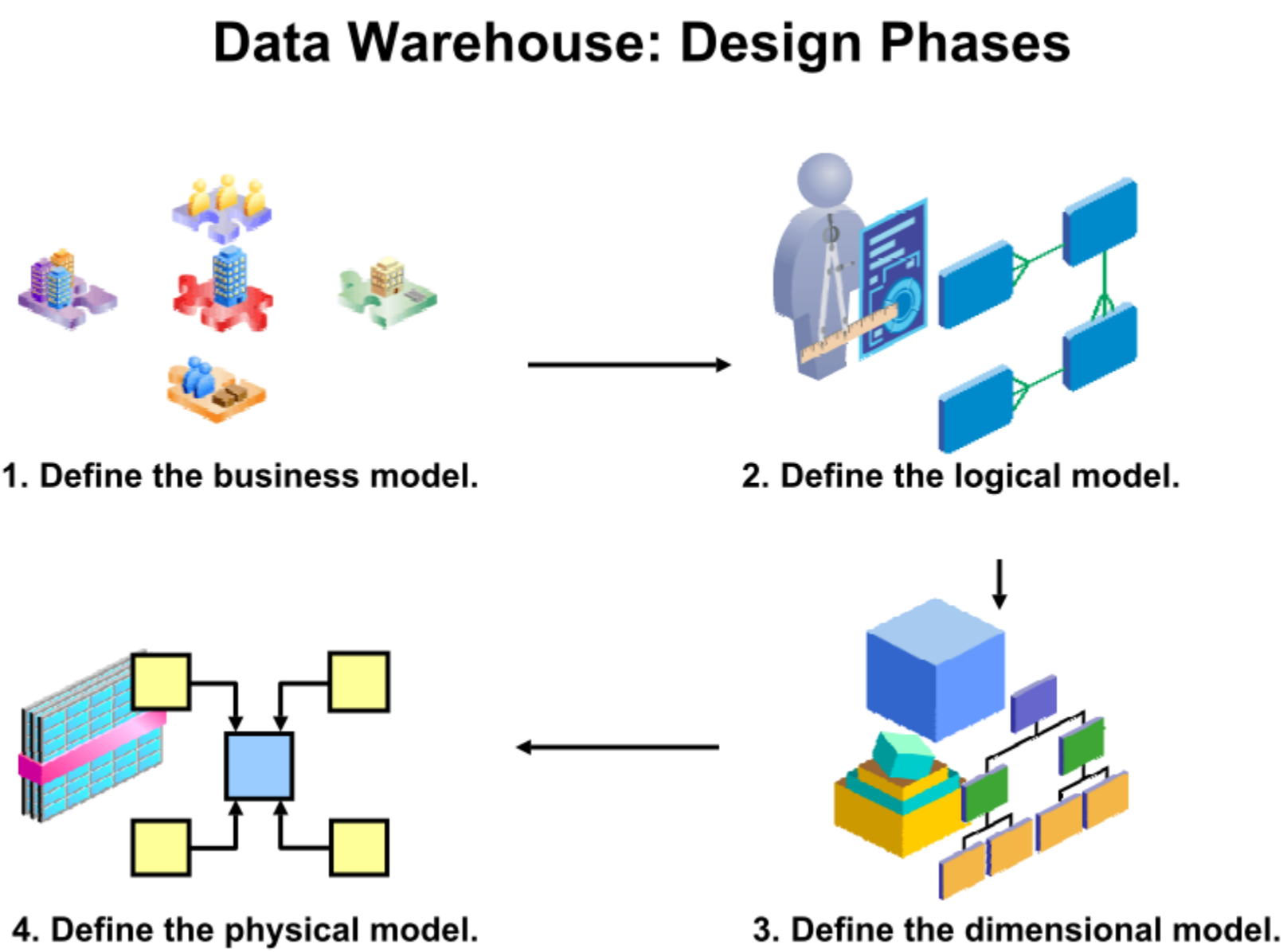
Źródło: M. Gorawski (2003) *Ocena efektywności architektur* … Op. cit., s. 2.

Porównanie architektur MOLAP i ROLAP

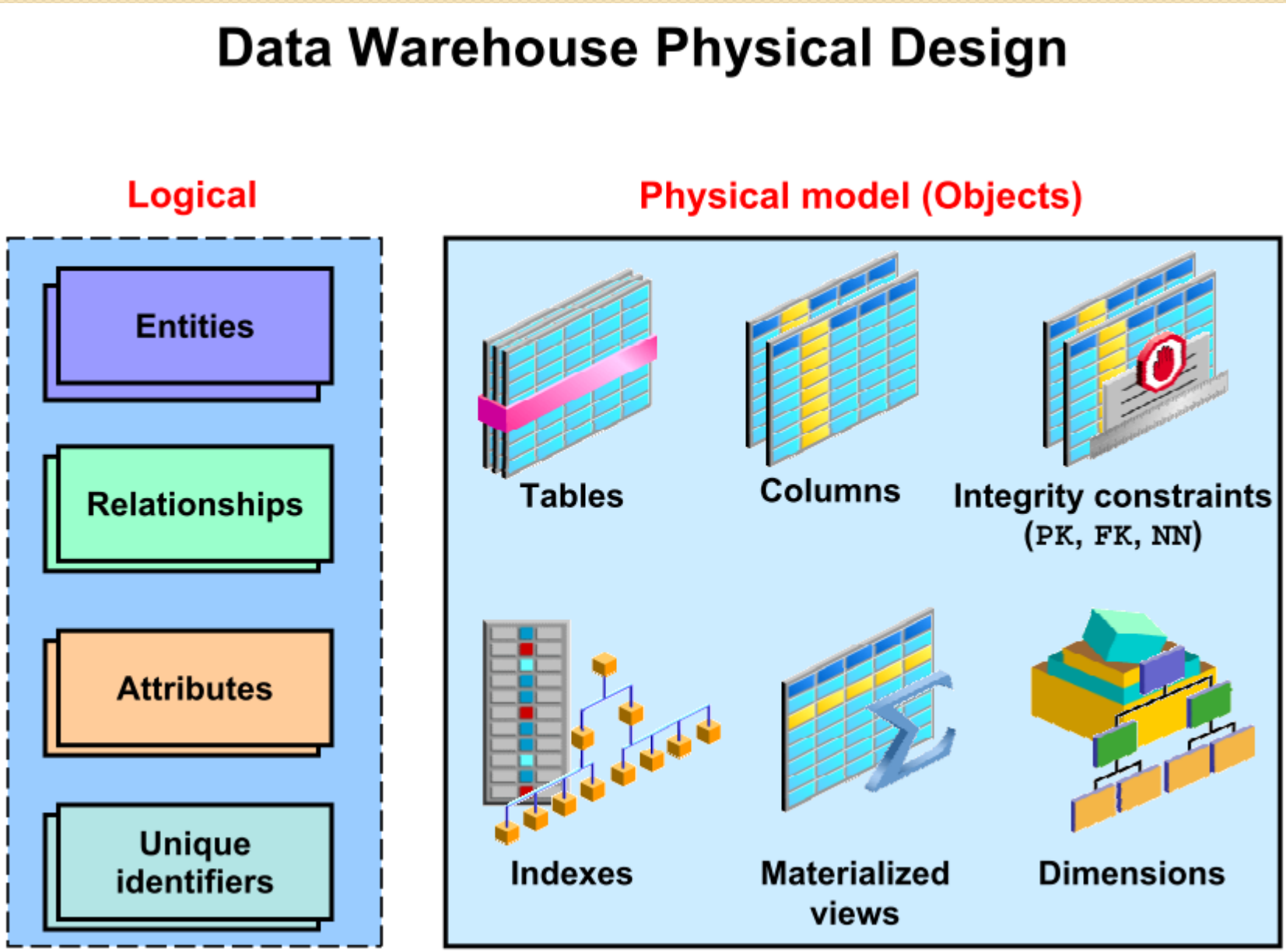
zgodnie z [2] (str. 60-62)

Architektura ROLAP daje znacznie większe możliwości

Modelowanie systemów ROLAP



Modelowanie systemów ROLAP



Modelowanie systemów ROLAP

TABELA FAKTÓW

- zawiera ilościowe (głównie liczbowe, ale mogą też być tekstowe) miary (*measures*), wskaźniki (*key figures*) cech istotnych dla danej działalności gospodarczej
- utrzymuje bardzo duże ilości danych
- szybko rozrasta się
- może zawierać dane podstawowe, wygenerowane oraz podsumowania
- typową własnością jest addytywność tych danych
- jest powiązana z tabelami wymiarów poprzez zdefiniowane w niej klucze obce, odnoszące się do kluczy głównych tabeli wymiarów, które zawierają cechy istotne dla danej działalności gospodarczej

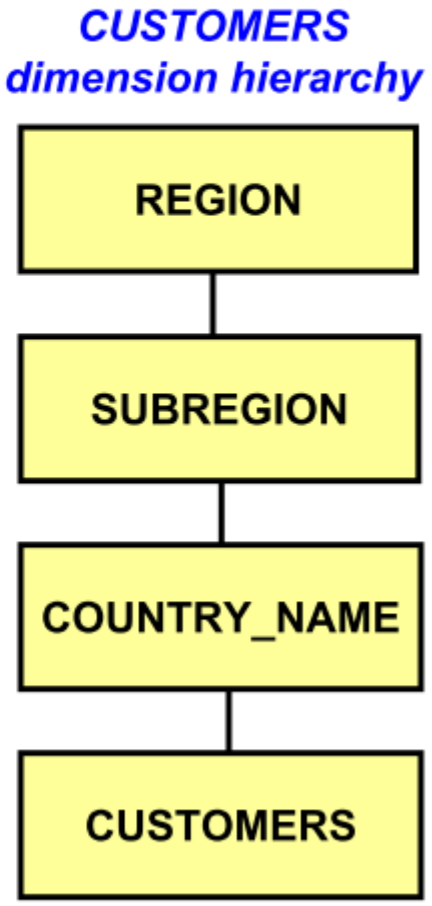
Sales (Fact Table)

```
PROD_ID
CUST_ID
TIME_ID
CHANNEL_ID
PROMO_ID
QUANTITY_SOLD
AMOUNT_SOLD
...
```

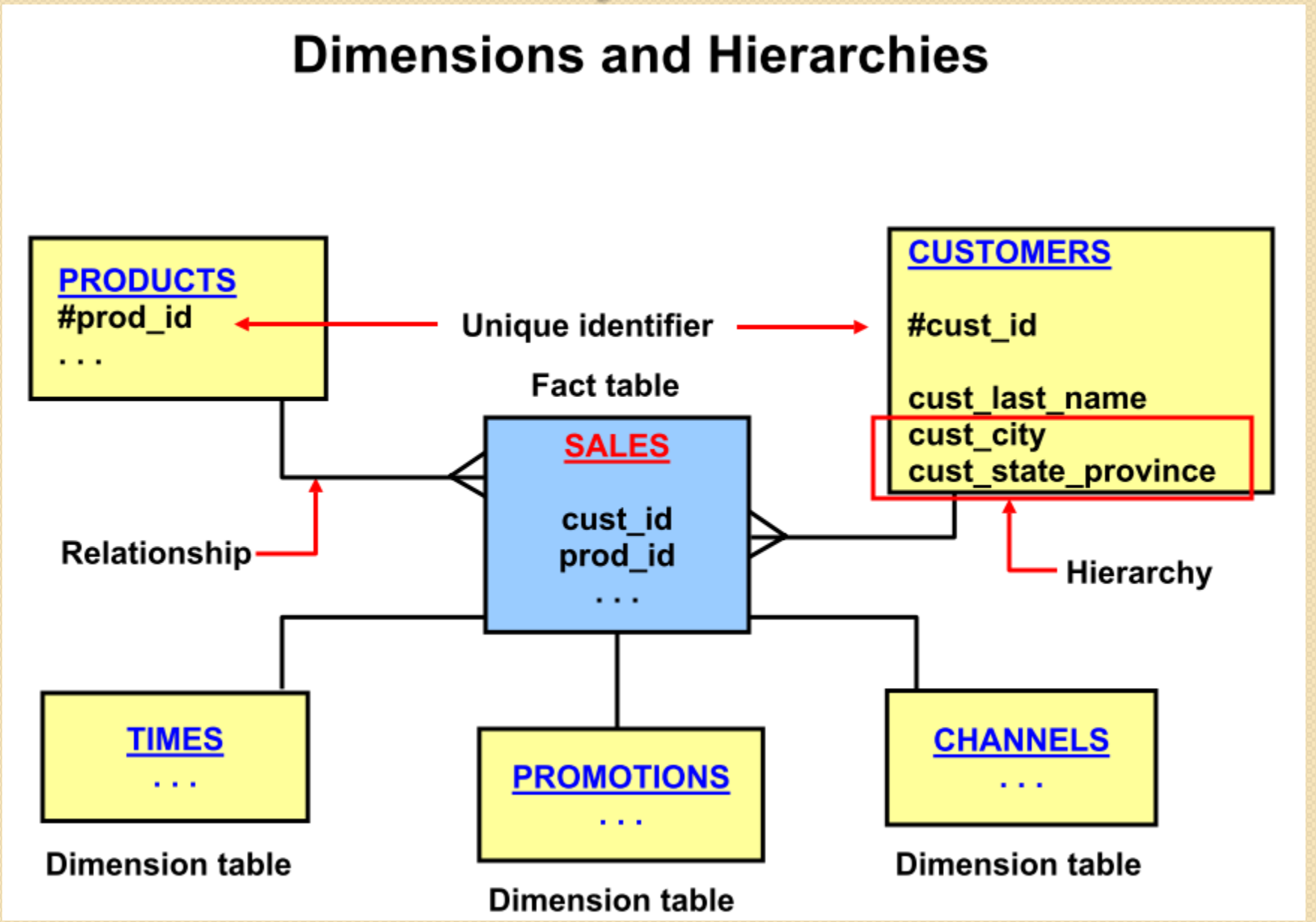
Modelowanie systemów ROLAP

WYMIARY I HIERARCHIE

- Wymiar składa się z cech istotnych dla danej działalności gospodarczej.
- Wymiar jest strukturą złożoną zwykle z jednej lub więcej hierarchii, które kategoryzują dane. Tzn. wymiar może mieć jedną lub więcej hierarchii, wspomagających jego wykorzystanie w analizie danych.
- Atrybuty wymiarów pomagają opisać wartości wymiarów
- Dane wymiarów są zwykle zebrane na najniższym poziomie szczegółowości i zagregowane do najwyższego poziomu ogólności
- Hierarchie:
  - porządkują poziomy, które organizują dane
  - umożliwiają agregowanie (zwijanie) oraz drążenie i rozwijanie danych



Modelowanie systemów ROLAP



Modelowanie systemów ROLAP

PODSTAWOWE SCHEMATY HURTOWNI DANYCH

- Organizacja obiektów w hurtowni danych jest przeprowadzana na różne sposoby:

  - schemat gwiazdy

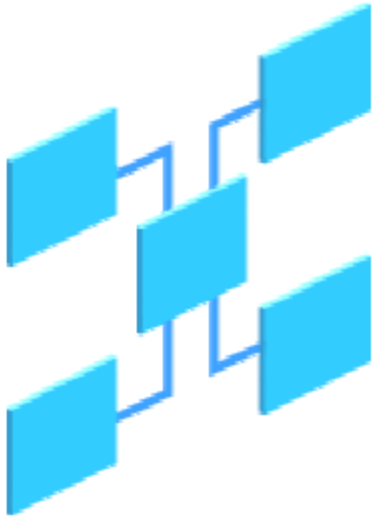
  - schemat płatka śniegu

  - schemat z zachowaną 3NF

  - schemat hybrydowy

- Przy wyborze schematu hurtowni danych należy kierować się modelem danych źródłowych i wymaganiami użytkownika

- Należy pamiętać, że w praktyce implementacja modelu logicznego na fizycznym systemie może wymagać zmian w zaprojektowanym modelu



Modelowanie systemów ROLAP

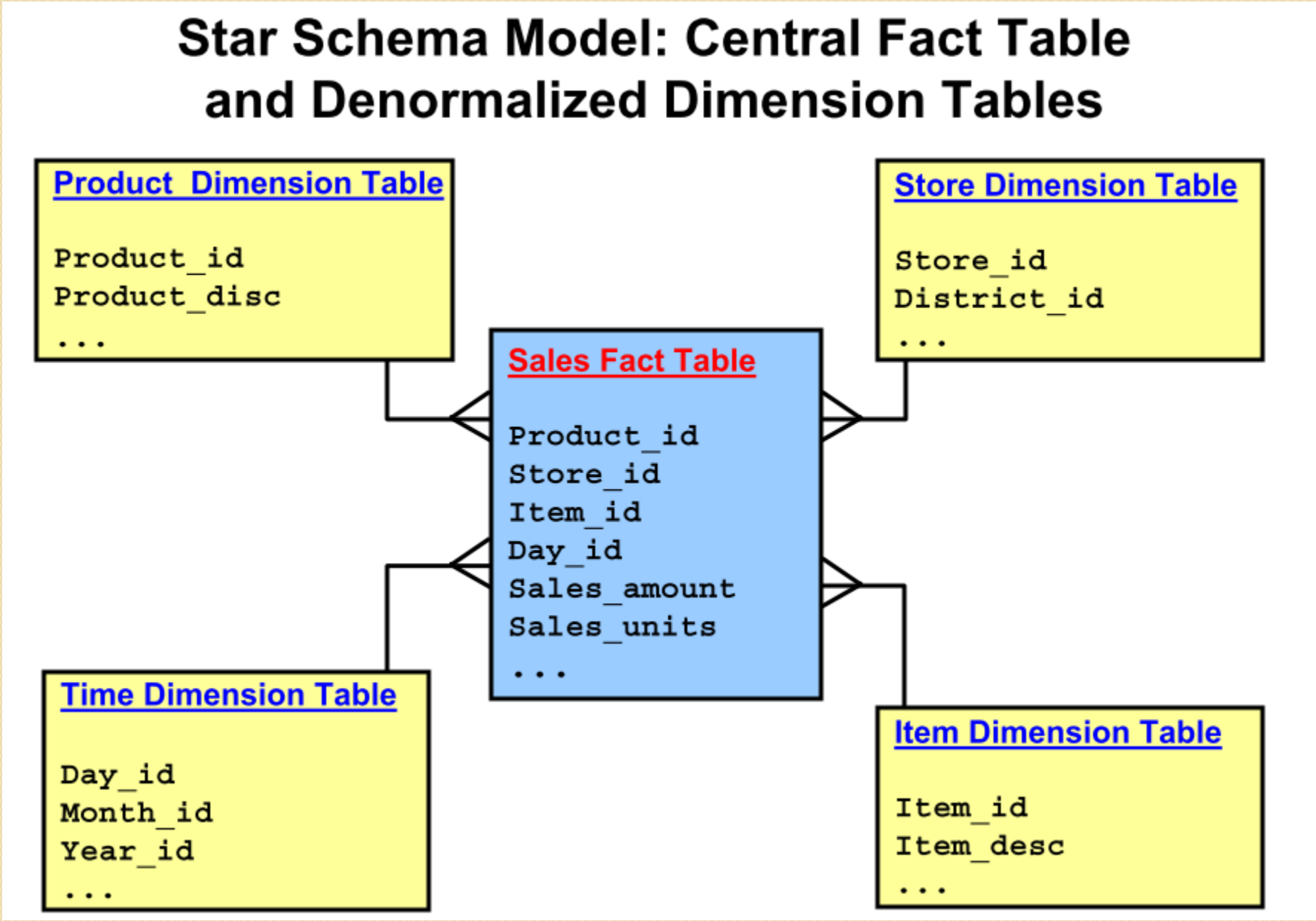
CECHY CHARAKTERYSTYCZNE SCHEMATÓW HURTOWNI DANYCH

- Schemat gwiazdy:
  - charakteryzuje go jedna lub więcej tabel faktów i pewna liczba dużo mniejszych tabel wymiarów
  - każda tabela wymiarów jest powiązana z tabelą faktów za pomocą związku klucz główny - klucz obcy

- Schemat płatka śniegu:
  - dane określonego wymiaru są pogrupowane w kilku mniejszych tabelach zamiast w jednej większej tabeli
  - ilość tabel wymiarów wzrasta, wymagając więcej powiązań za pomocą klucz obcych

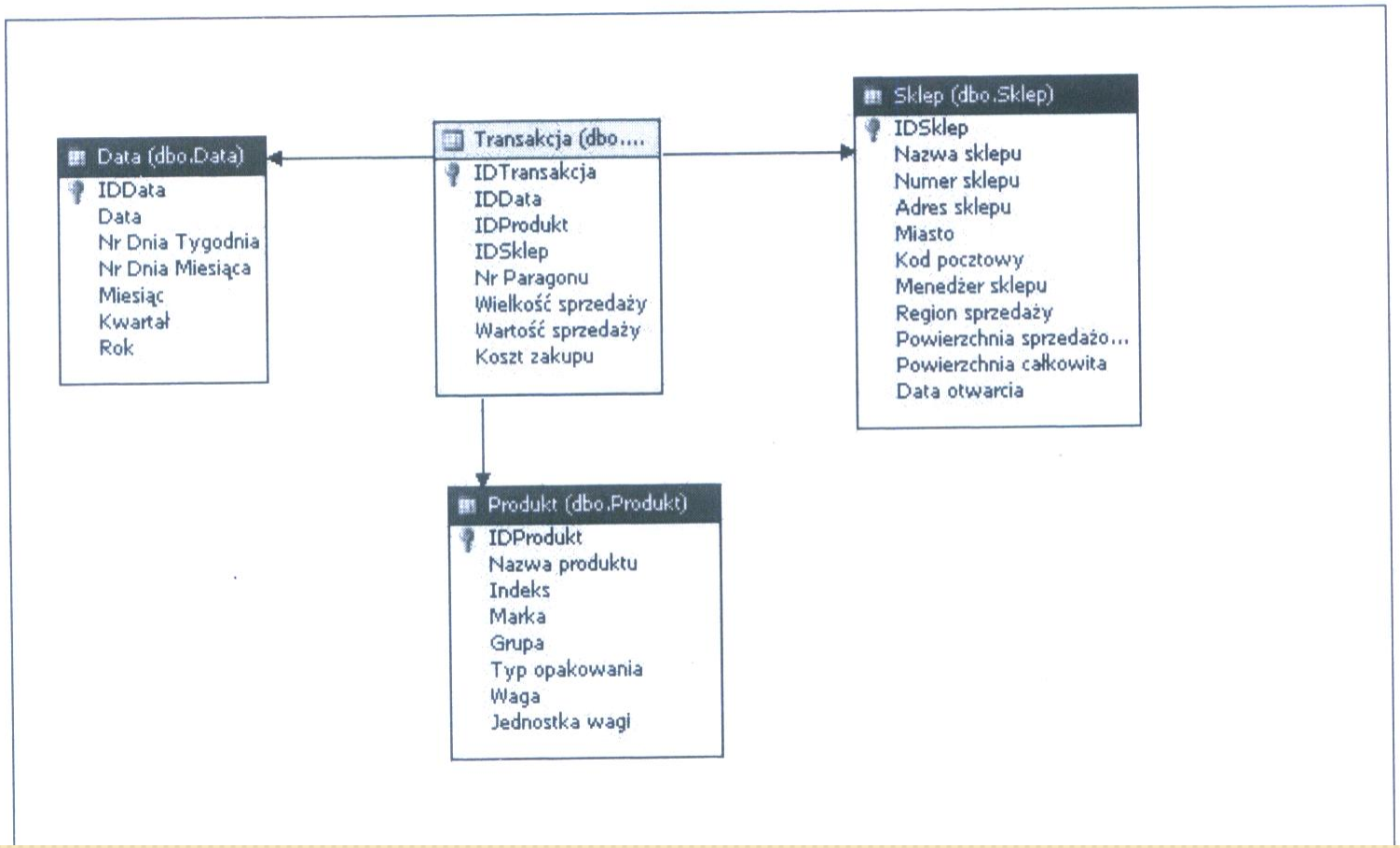
- Schemat z zachowaną 3NF:
  - jest to klasyczny model relacyjnej bazy danych, w którym redundancja danych jest zminimalizowana dzięki normalizacji

Modelowanie systemów ROLAP



Modelowanie systemów ROLAP

Zdefiniowano następującą tablicę faktów Transakcja (rysunek zawiera też wyspecyfikowane wcześniej wymiary – schemat gwiaździsty):



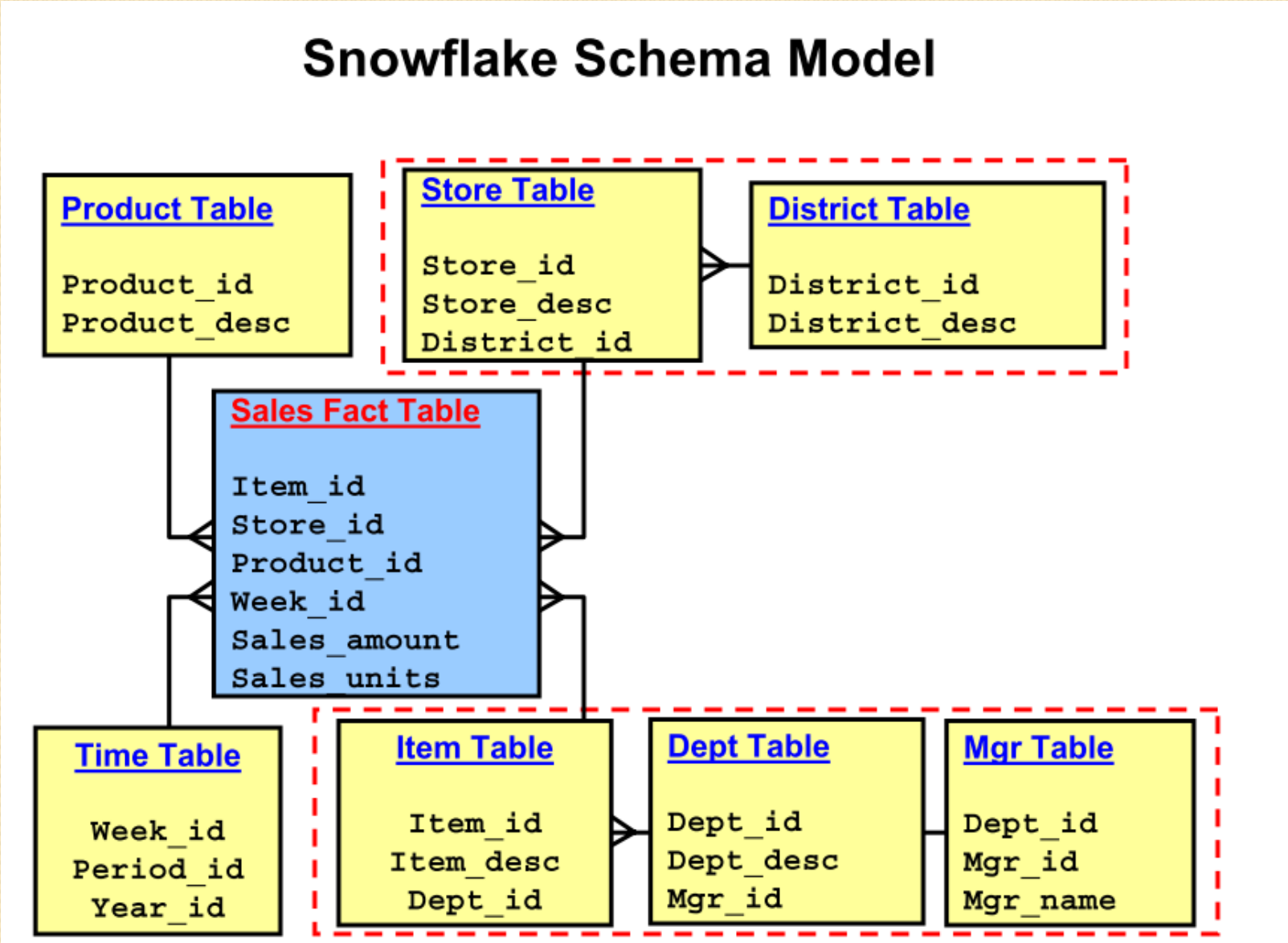
Przykład schematu gwiaździstego [1] (str. 34)

Modelowanie systemów ROLAP

Zalety schematu gwiazdy:

- wspiera wielowymiarową analizę
- jest modelem, który poprawia wydajność hurtowni danych
- umożliwia optymalizatorowi przygotowanie lepszych planów wykonania zapytań
- jest modelem łatwo rozszerzalnym
- poszerza możliwości wyboru narzędzi dostępu do danych

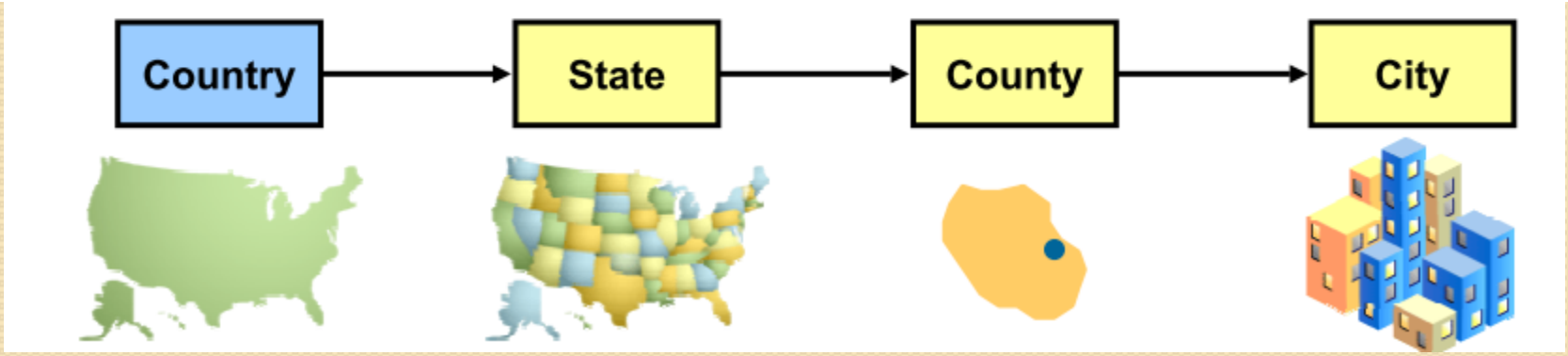
Modelowanie systemów ROLAP



Modelowanie systemów ROLAP

Wady i zalety schematu płatka śniegu:

- Może być bezpośrednio obsługiwany tylko przez wybrane narzędzia.
- Jednak jest modelem, który łatwiej zmieniać i umożliwia szybkie ładowanie danych.
- Niestety, taka hurtownia może stać się zbyt duża i nie będziemy w stanie nią zarządzać
- Ponadto wydajność zapytań jest w nim niższa a metadane są bardziej złożone



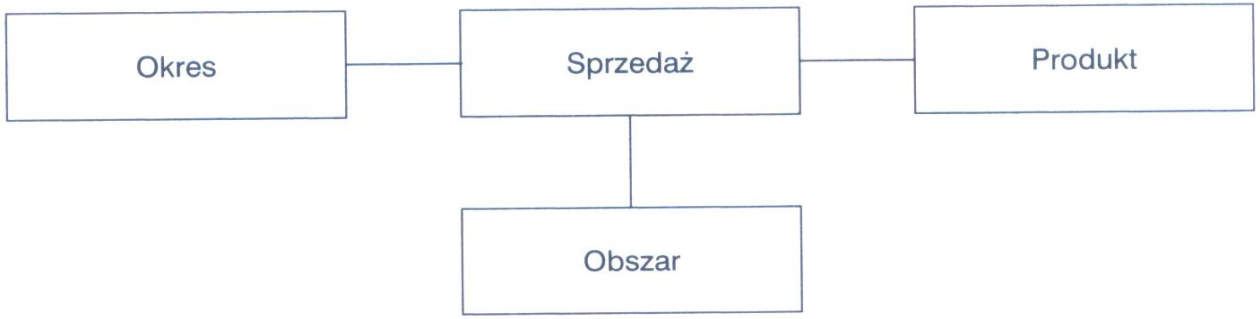
Podstawowe operacje analizy danych wielowymiarowych

- określenie zakresu analizy
- drążenie-rozwijanie (*drill down*)
- zwijanie (*roll up*)
- wycinanie (selekcja, *slice and dice*)
- obracanie (rotating)
- działania na faktach (wskaźnikach liczbowych)
  - działania arytmetyczne i funkcje agregujące
  - ranking, sortowanie z podziałem na grupy
  - obliczenie wskaźników ekonomicznych
  - stosowanie różnych modeli statystycznych i ekonometrycznych

Podstawowe operacje analizy danych wielowymiarowych

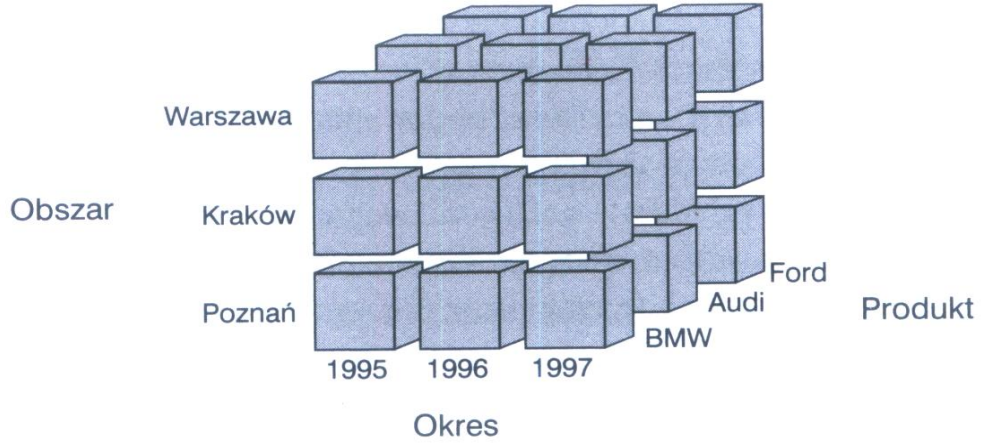
OKREŚLENIE ZAKRESU ANALIZY

Konstrukcja przykładowej kostki wielowymiarowej opartej
na hurtowni danych zorganizowanej zgodnie ze schematem gwiazdy [1] (str. 38)



Rysunek 3.1. Przykładowy model wielowymiarowy dla sieci komisów samochodowych

Źródło: opracowanie własne.



Rysunek 3.2. Kostka wielowymiarowa dla sprzedaży w komisie samochodowym

Źródło: Wrembel i in. 2004.

Podstawowe operacje analizy danych wielowymiarowych

Przykładowe raporty [1] (str. 39) dla skonstruowanej kostki wielowymiarowej opartej na hurtowni danych zorganizowanej zgodnie ze schematem gwiazdy

Tabela 3.1. Przykładowy raport ze sprzedaży w komisie samochodowym (wersja 1)

Okres	Obszar	Sprzedaż ilościowo
1995	Warszawa	1000
	Kraków	500
	Poznań	1000
1996	Warszawa	1500
	Kraków	500
	Poznań	900
1997	Warszawa	2000
	Kraków	500
	Poznań	800

Źródło: opracowanie własne.

Tabela 3.2. Przykładowy raport ze sprzedaży w komisie samochodowym (wersja 2)

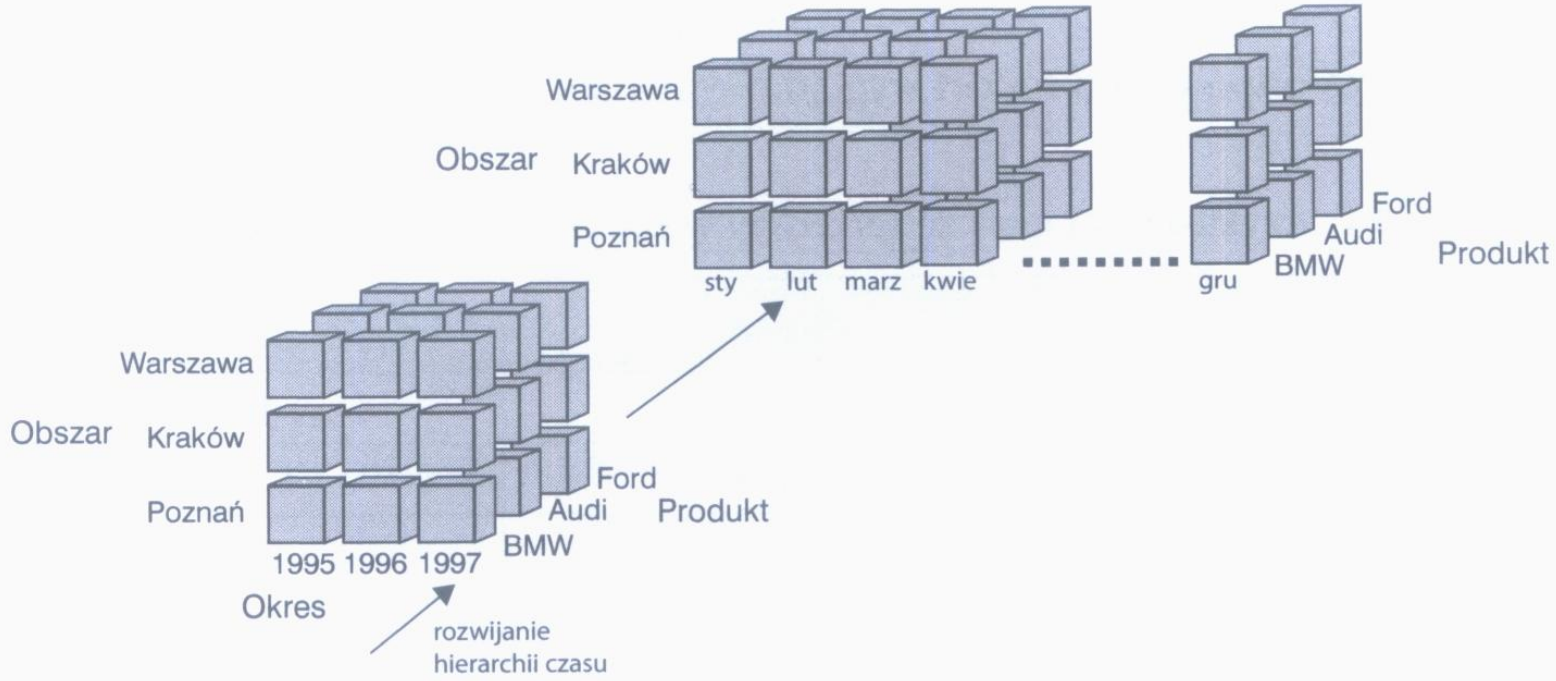
Sprzedaż ilościowo dla:

Okres	Obszar		
	Warszawa	Kraków	Poznań
1995	1000	500	1000
1996	1500	500	900
1997	2000	500	800

Źródło: opracowanie własne.

Podstawowe operacje analizy danych wielowymiarowych

DRĄŻĘNIE-ROZWIJANIE danych
w skonstruowanej kostki wielowymiarowej opartej
na hurtowni danych zorganizowanej zgodnie ze schematem gwiazdy [1] (str. 40)



Rysunek 3.3. Operacja drążenia w wymiarze Okres

Źródło: Wrembel i in. 2004.

ZWIJANIE (AGREGOWNIE) danych jest operacją odwrotną do drążenia-rozwijania

Podstawowe operacje analizy danych wielowymiarowych

Przykładowy raport wynikający z drążenia-rozwijania hierarchii czasu dla wymiaru OKRES skonstruowanej kostki wielowymiarowej [1] (str. 40 i 41)

Tabela 3.3. Przykładowy raport ze sprzedaży w komisie samochodowym po drążeniu w wymiarze: Okres

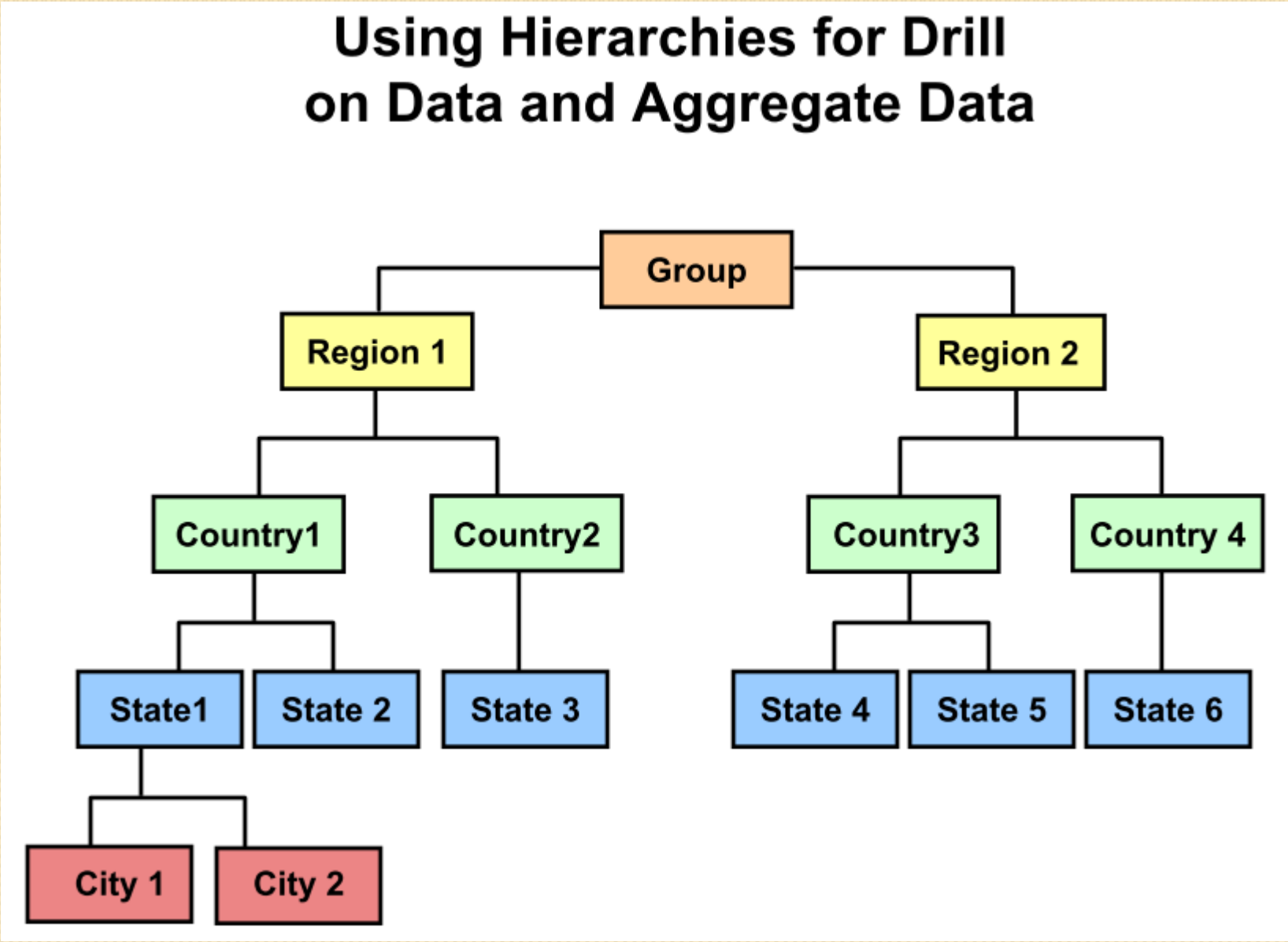
Sprzedaż ilościowo dla:

Okres	Obszar		
	Warszawa	Kraków	Poznań
01/1997	100	30	30
02/1997	120	20	20
03/1997	150	20	20
04/1997	200	30	30
05/1997	100	40	90
06/1997	190	50	100
07/1997	230	60	120
08/1997	200	70	100
09/1997	220	70	120
10/1997	220	30	60
11/1997	180	40	70
12/1997	90	40	40

Źródło: opracowanie własne.

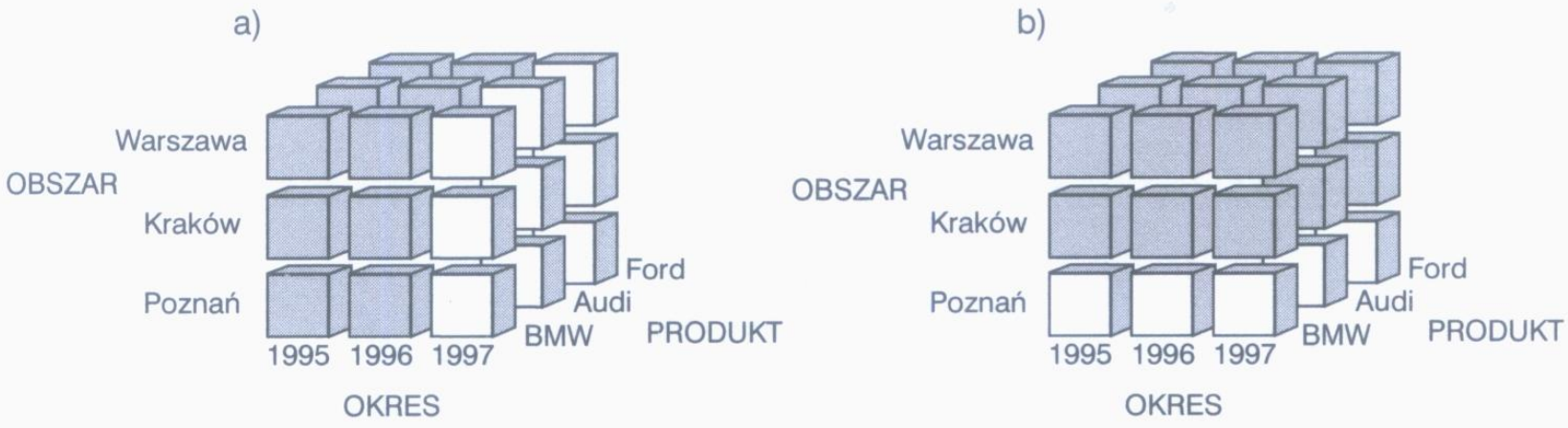
Podstawowe operacje analizy danych wielowymiarowych

Schemat zastosowania hierarchii obszaru do DRĄŻĘNIA-ROZWIJANIA i ZWIJANIA (AGREGOWANIA) danych



Podstawowe operacje analizy danych wielowymiarowych

WYCINANIE (SELEKCJA)
Konstrukcja przykładowej kostki wielowymiarowej opartej
na hurtowni danych zorganizowanej zgodnie ze schematem gwiazdy [1] (str. 41)



Rysunek 3.4. Operacja wycinania a) wymiar Okres = 1997; b) wymiar Obszar = Poznań

Źródło: Wrembel i in. 2004.

Podstawowe operacje analizy danych wielowymiarowych

Przykładowe raporty dla wykonanych selekcji na skonstruowanej kostce wielowymiarowej [1] (str. 42)

Tabela 3.4. Przykładowy raport ze sprzedaży w komisie samochodowym po wycięciu w wymiarze: Okres

Sprzedaż ilościowo dla: Okres = 1997:

Produkt	Obszar		
	Warszawa	Kraków	Poznań
BMW	1000	150	300
Audi	500	250	300
Ford	500	100	200

Źródło: opracowanie własne.

Tabela 3.5. Przykładowy raport ze sprzedaży w komisie samochodowym po wycięciu w wymiarze: Obszar

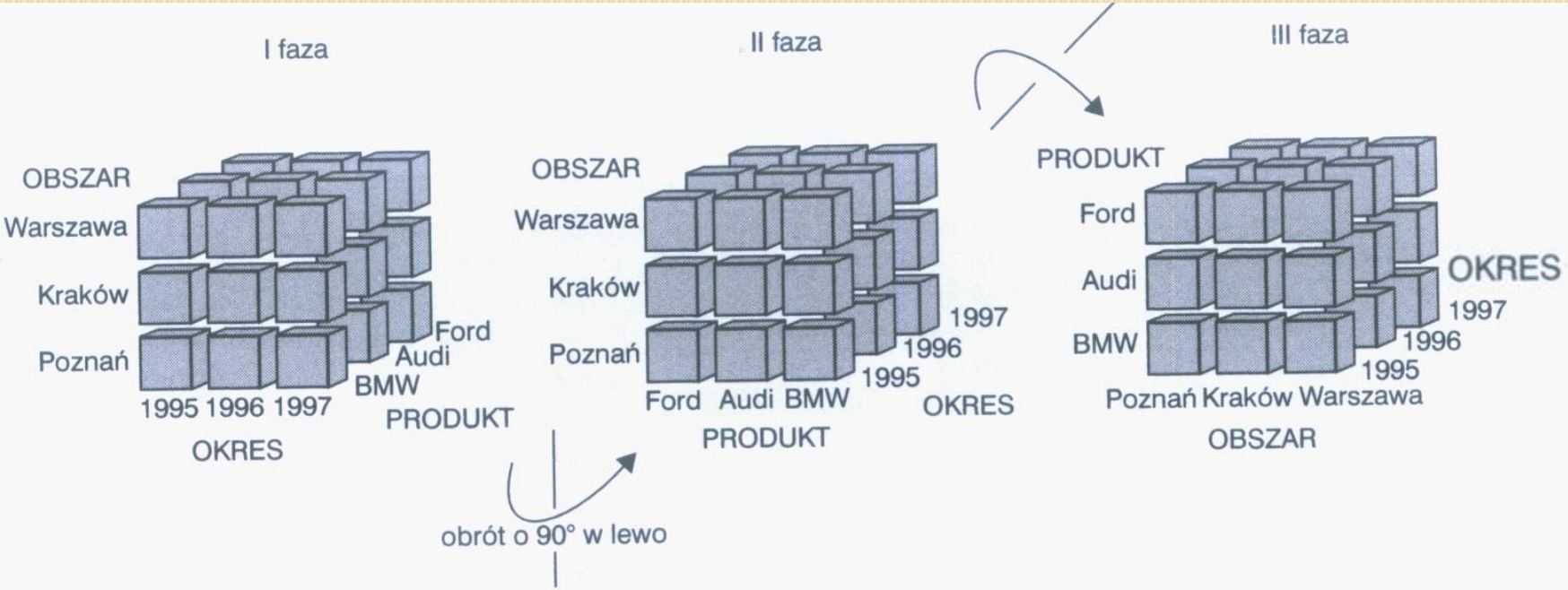
Sprzedaż ilościowo dla: Obszar = Poznań:

Produkt	Okres		
	1995	1996	1997
BMW	200	300	300
Audi	300	300	300
Ford	500	300	200

Źródło: opracowanie własne.

Podstawowe operacje analizy danych wielowymiarowych

OBRACANIE (patrzenie na kostkę z danymi z różnych punktów widzenia) [1] (str. 43)



Rysunek 3.5. Operacja obracania kostki wielowymiarowej

Źródło: Wrembel i in. 2004.

Podstawowe operacje analizy danych wielowymiarowych

Przykładowe raporty dla poszczególnych faz dokonanego obrotu naszej wielowymiarowej kostki [1] (str. 43)

Tabela 3.6. Przykładowy raport ze sprzedaży w komisie samochodowym w I fazie operacji obracania

Sprzedaż ilościowo:

Okres	Obszar		
	Warszawa	Kraków	Poznań
1995	1000	500	1000
1996	1500	500	900
1997	2000	500	800

Źródło: opracowanie własne.

Podstawowe operacje analizy danych wielowymiarowych

Przykładowe raporty dla poszczególnych faz dokonanego obrotu naszej wielowymiarowej kostki [1] (str. 43)

Tabela 3.7. Przykładowy raport ze sprzedaży w komisie samochodowym w II fazie operacji obracania

Sprzedaż ilościowo:

Produkt	Obszar		
	Warszawa	Kraków	Poznań
BMW	1000	500	800
Audi	1500	500	900
Ford	2000	500	1000

Źródło: opracowanie własne.

Podstawowe operacje analizy danych wielowymiarowych

Przykładowe raporty dla poszczególnych faz dokonanego obrotu naszej wielowymiarowej kostki [1] (str. 44)

Tabela 3.8. Przykładowy raport ze sprzedaży w komisie samochodowym w III fazie operacji obracania

Sprzedaż ilościowo:

Obszar	Produkt		
	BMW	Audi	Ford
Warszawa	1000	1500	2000
Kraków	500	500	500
Poznań	800	900	1000

Źródło: opracowanie własne.