

Hurtownie danych – 1

Problematyka hurtowni danych

Wykład przygotował:

Robert Wrembel

ZSBD – wykład 12 (1)



Plan wykładu

- Problematyka integracji danych
- Integracja danych za pomocą hurtowni danych
- Przetwarzanie analityczne OLAP
- Model wielowymiarowy
- Implementacje modelu wielowymiarowego
 - ROLAP
 - MOLAP

ZSBD – wykład 12 (2)

Celem niniejszego wykładu jest wprowadzenie do zagadnień integracji danych i technologii hurtowni danych. W ramach wykładu zostaną omówione następujące zagadnienia:

- wprowadzenie do problematyki integracji rozproszonych i heterogenicznych baz danych,
- podstawowa architektura integracji danych oparta o hurtownię danych,
- wprowadzenie i charakterystyka przetwarzania analitycznego (On-Line Analytical Processing - OLAP),
- wielowymiarowy model danych,
- implementacja modelu wielowymiarowego w serwerach relacyjnych (ROLAP) i wielowymiarowych (MOLAP).



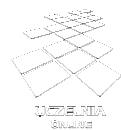
Cechy systemów informatycznych

- Składowanie, przetwarzanie i zapewnienie efektywnego dostępu do danych w przedsiębiorstwie
- Akutalny stan technologiczny systemów informatycznych
 - heterogeniczność
 - rozproszenie
- Heterogeniczność
 - wielość struktur danych
 - różna funkcjonalność
 - różne modele danych
- Rozproszenie
 - dane są rozmieszczone w geograficznie różnych lokalizacjach

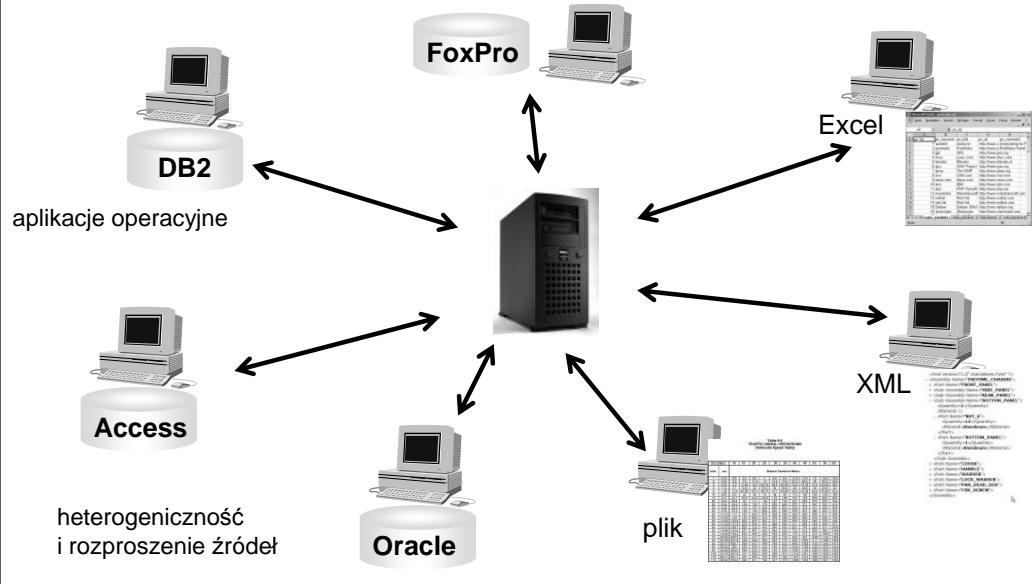
ZSBD – wykład 12 (3)

Główym zadaniem systemów informatycznych jest przyspieszenie i ułatwienie pracy poprzez efektywny dostęp, przetwarzanie i składowanie danych. Dane gromadzone przez firmy i instytucje są często przechowywane w heterogenicznych i rozproszonych systemach informatycznych.

Heterogeniczność oznacza, że systemy posiadają różne struktury, funkcjonalność i wykorzystują różne modele danych (np. hierarchiczne, relacyjne, obiektowe, semistrukturalne), przechowują dane w dokumentach tekstowych, czy arkuszach kalkulacyjnych. Często nawet w ramach tej samej instytucji wykorzystuje się różne systemy informatyczne. Dodatkowym problemem w dostępie do informacji jest geograficzne rozproszenie źródeł danych.



Problematyka integracji danych



ZSBD – wykład 12 (4)

Problematykę integracji danych ilustruje niniejszy slajd.



SI we wspomaganiu zarządzaniem

- Wspomaganie zarządzaniem bazują na analizie danych
- Dane opisują bieżący stan i historię działania firmy
- Analizie powinny podlegać możliwie wszystkie dane opisujące analizowany fragment działalności firmy
- Problem
 - dane są składowane w systemach heterogenicznych
 - dane są składowane w systemach rozproszonych
 - nieefektywny dostęp i analiza danych
- Rozwiązanie
 - integracja danych na platformie hurtowni danych

ZSBD – wykład 12 (5)

Racjonalizacja działania firm i instytucji wymaga stosowania procesów wspomagania decyzji kadry zarządzającej. Procesy wspomagania decyzji bazują na danych analitycznych opisujących bieżący stan i historię działania danej firmy. Programowe narzędzia analityczne, na podstawie danych elementarnych będących wynikiem pracy pojedynczych pracowników, powinny udostępniać informacje statystyczne o bieżącym stanie firmy, występujących trendach i korelacjach między różnymi czynnikami. Dzięki szybkiej analizie bazującej na pełnej i aktualnej informacji o stanie firmy, kadra zarządzająca może podejmować trafniejsze decyzje o strategicznym znaczeniu dla rozwoju danego przedsiębiorstwa.

W celu zapewnienia decydentom pełnego dostępu do heterogenicznych i rozproszonych danych, buduje się systemy integrujące. Jedną z najczęściej stosowanych technik integracji danych jest ich transformacja do wspólnego modelu i składowanie w centralnym systemie, zwany hurtownią (magazynem) danych (ang. data warehouse).



Hurtownia danych

- Bardzo duża baza danych (dziesiątki, setki TB)
- Charakterystyka:
 - dane są wyłącznie odczytywane przez użytkowników
 - zawiera dane historyczne i bieżące
 - zawiera dane zagregowane na wielu poziomach
 - zawartość jest zorientowana tematycznie

ZSBD – wykład 12 (6)

Hurtownia danych jest bardzo dużą bazą danych (często rzędu dziesiątek czy setek terabajtów). Hurtownia danych posiada następującą charakterystykę:

- dane w niej składowane nie są modyfikowane przez użytkowników, są natomiast wyłącznie przez nich odczytywane,
- zawiera wszystkie dane historyczne i bieżące,
- zawiera dane zagregowane na wielu poziomach szczegółowości,
- zawartość hurtowni jest zorientowana tematycznie, np. hurtownia danych nt. ruchu telefonicznego i opłat klientów sieci komórkowej, hurtownia danych nt. sprzedaży pojazdów.



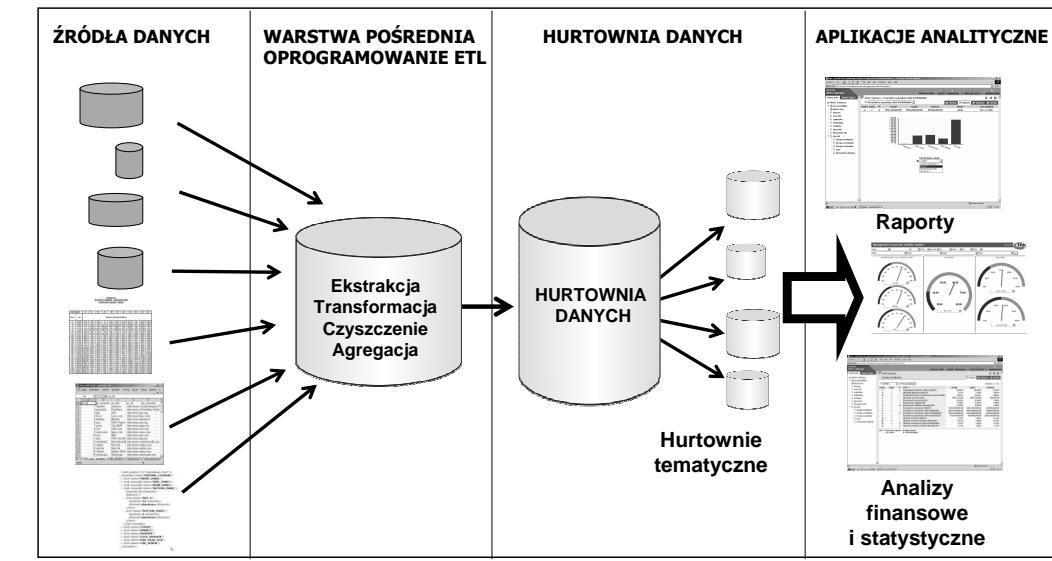
Systemy komercyjne

- Oracle8i, Oracle9i, Oracle10g – Oracle Corporation,
- DB2 UDB – IBM,
- SybaseIQ – Sybase, Inc.,
- SAS Enterprise BI Server - SAS Institute
- MS SQL Server – Microsoft,
- SAP Business Warehouse – SAP,
- Adabas C i Adabas D – Software AG,
- Teradata – NCR Corporation,
- Hyperion Essbase OLAP Server – Hyperion Solutions Corporation
- Red Brick Warehouse – Red Brick Systems

ZSBD – wykład 12 (7)



Podstawowa architektura HD

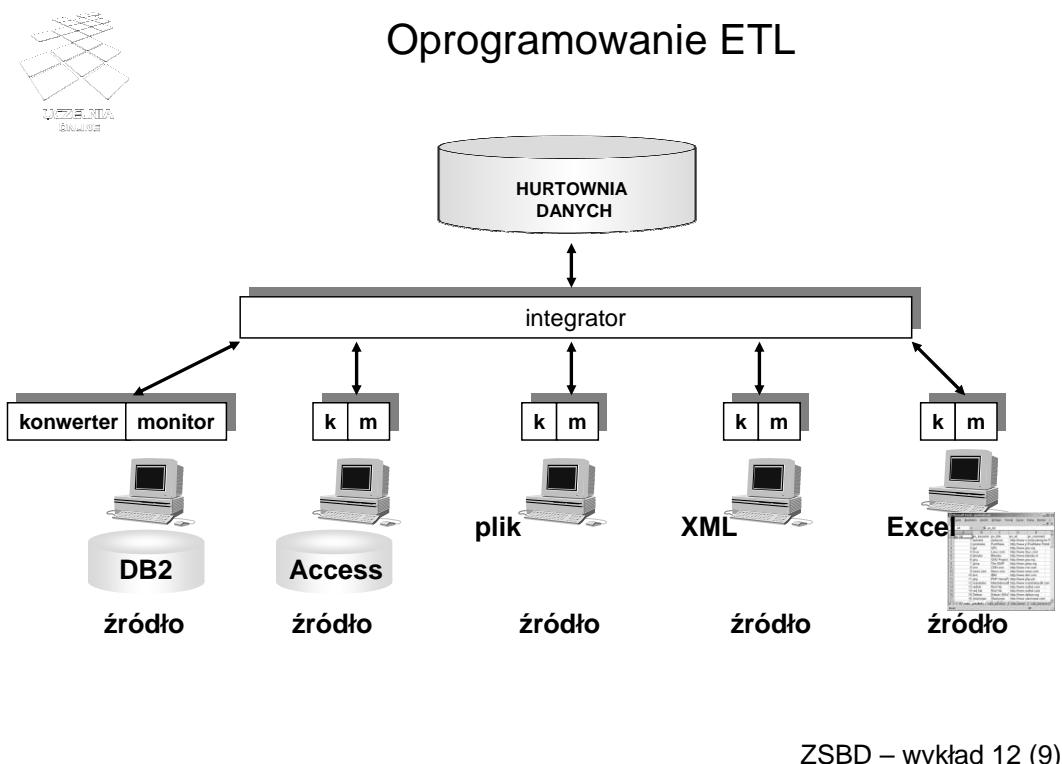


ZSBD – wykład 12 (8)

Podstawową poglądową architekturę hurtowni danych przedstawiono na slajdzie. Heterogeniczne i rozproszone źródła danych zasilają hurtownię danymi za pośrednictwem warstwy oprogramowania ETL. Jego podstawowymi zadaniami są wykrywanie zmian w źródłach, transformacja danych do wspólnej postaci, uspójnianie i czyszczenie danych, agregowanie danych. Uspójnione dane są następnie ładowane do centralnej hurtowni danych.

Centralna hurtownia danych najczęściej zawiera dane dla wszystkich grup decydentów. Ze względu na ilość danych, wygodniej jest zbudować na bazie hurtowni centralnej, małe hurtownie tematyczne (ang. data marts), zawierające dane opisujące wąskie dziedziny funkcjonowania firmy. Taka hurtownia tematyczna zawiera często dane na jeszcze wyższym poziomie agregacji niż hurtownia centralna. Przykładowo, hurtownia danych nt. ruchu telefonicznego i opłat klientów sieci komórkowej może posłużyć do zbudowania dwóch data marts opisujących odpowiednio natężenie ruchu telefonicznego w ciągu dnia i ranking klientów ze względu na płacone rachunki.

Na hurtowni danych pracują tzw. aplikacje analityczne wspomagania decyzji (ang. decision support), czy odkrywania wiedzy (ang. data mining), których celem jest wspomaganie pracy kadry zarządzającej poprzez analizę trendów, anomalii, poszukiwanie reguł zachowań.



ZSBD – wykład 12 (9)

Oprogramowanie ETL (Extraction Translation Loading) realizuje tzw. procesy ETL, składające się z trzech następujących faz:

- odczytu danych ze źródeł (Extraction),
- transformacji ich do wspólnego modelu wykorzystywanego w magazynie wraz z usunięciem wszelkich niespójności (Translation),
- wczytanie danych do magazynu (Loading). Na slajdzie przedstawiono podstawowe komponenty oprogramowania ETL.

Obiekty oznaczone jako źródło reprezentują heterogeniczne i rozproszone źródła danych. Z każdym z takich źródeł jest związana dedykowana dla niego warstwa oprogramowania o nazwie konwerter/monitor.

Zadaniem modułu konwertera jest transformowanie danych z formatu wykorzystywanego w źródle, do formatu wykorzystywanego w hurtowni. Dlatego, dla każdego modelu danych źródłowych konieczne jest zastosowanie specyficznego modułu konwertera. Przykładowo, jeśli źródło przechowuje dane w dokumentach tekstowych, a hurtownia została zaprojektowana z wykorzystaniem modelu relacyjnego, to konwerter musi zapewnić poprawne odwzorowanie danych z plików w strukturę modelu relacyjnego.

Zadaniem modułu monitora jest wykrywanie zmian w danych źródłowych i ich przekazywanie do warstwy oprogramowania integratora (po uprzedniej konwersji do modelu danych hurtowni). Sposób wykrywania zmian w danych źródłowych zależy od własności samych źródeł.



Wykrywanie zmian w źródłach

- Źródła aktywne
- Źródła utrzymujące dzienniki operacji
- Źródła przepytywalne
- Źródła wspierające migawki

ZSBD – wykład 12 (10)

Wyróżnia się cztery następujące rodzaje źródeł danych:

- aktywne (ang. active sources),
- utrzymujące dzienniki operacji wykonywanych na danych źródłowych (ang. logged sources),
- przepytywalne (ang. queryable sources),
- wspierające mechanizm migawek (ang. snapshot sources).

Źródła aktywne posiadają zaimplementowane mechanizmy wyzwalaczy, które informują monitor o zmianach zachodzących w danych źródłowych.

W źródłach utrzymujących dzienniki operacji zmiany są wykrywane poprzez analizę zawartości dziennika w module monitora.

Źródła przepytywalne umożliwiają wydawanie zapytań i w celu wykrycia zmian w danych źródłowych monitor okresowo wydaje zapytania do takich źródeł.

Źródła wspierające mechanizm migawek umożliwiają tworzenie migawek, czyli obrazów stanu źródła z określonego momentu. Porównanie migawek z kolejnych momentów umożliwia wykrycie zmian.



Przetwarzanie analityczne OLAP (1)

- Aplikacje analityczne
 - wspomagania decyzji
 - eksploracji danych
- Zorientowane na wspieranie procesów decyzyjnych
 - wykonywanie zaawansowanych analiz, wspomagających zarządzanie przedsiębiorstwem, np.
 - analiza trendów sprzedaży
 - analiza nakładów reklamowych i zysków
 - analiza ruchu telefonicznego

ZSBD – wykład 12 (11)

Na hurtowni danych pracują tzw. aplikacje analityczne wspomagania decyzji (ang. decision support), czy eksploracji danych (ang. data mining). Aplikacje analityczne (ang. On-Line Analytical Processing - OLAP) są zorientowane na wspieranie procesów decyzyjnych, czyli przetwarzanie danych historycznych i zagregowanych. Przykładami takich zapytań mogą być: *Jaki jest trend sprzedaży towarów z branży AGD w ostatnich kilku tygodniach? Jaki jest rozkład sprzedaży lodówek w województwie wielkopolskim?*



Przetwarzanie analityczne OLAP (2)

- Operacje
 - łączenia (kilka, kilkanaście, kilkadziesiąt tabel),
 - filtrowania,
 - agregowania (np. suma, średnia)
- Dostęp do ogromnych wolumenów danych (miliony, dziesiątki, setki milionów rekordów)
- Czas realizacji zapytań analitycznych: godziny, dziesiątki godzin
- Problem efektywności, czasu odpowiedzi na zapytanie OLAP

ZSBD – wykład 12 (12)

Jak wspomniano, większość operacji realizowanych przez aplikacje analityczne obejmuje złożone zapytania wykorzystujące łączenie wielu tabel, filtrowanie, agregowanie, operujące na milionach rekordów. W konsekwencji, czasy przetwarzania danych sięgają nawet dziesiątek godzin. Kluczowym parametrem efektywności tego typu systemów jest czas odpowiedzi na zapytania analityczne.



Wielowymiarowy model danych (1)

- Dane zorganizowane w postaci wielowymiarowego modelu danych (ang. Multidimensional Data Model)
 - fakty
 - wymiary
- Fakty
 - informacje podlegające analizie
 - sprzedaż, rozmowy telefoniczne, ubezpieczenia
 - charakteryzowane ilościowo za pomocą miar
 - liczba sprzedanych sztuk towaru, czas trwania rozmowy, kwota ubezpieczenia

ZSBD – wykład 12 (13)

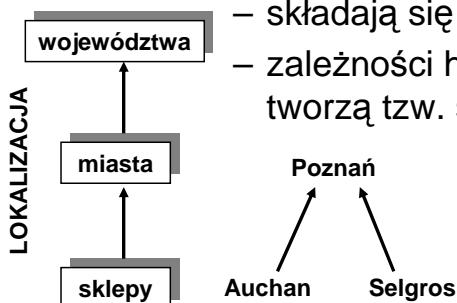
Dane w magazynie są zorganizowane w postaci tzw. modelu wielowymiarowego (ang. multidimensional data model), w którym wyróżnia się dwie podstawowe kategorie danych, tj. fakty i wymiary.

Fakty (ang. facts) reprezentują informacje podlegające analizie, np. fakt sprzedaży produktu, fakt wykonania rozmowy telefonicznej, fakt ubezpieczenia pojazdu. Fakty są charakteryzowane ilościowo za pomocą cech zwanych **miarami** (ang. measures). Przykładowo, miarą jest liczba zakupionych produktów, czas trwania rozmowy, kwota ubezpieczenia.



Wielowymiarowy model danych (2)

- Wymiary
 - ustalają kontekst analizy
 - sprzedaż czekolady (produkt) w Auchan (sklep) w poszczególnych miesiącach roku (czas)
 - składają się z poziomów, które tworzą hierarchię
 - zależności hierarchiczne między poziomami tworzą tzw. strukturę wymiaru



ZSBD – wykład 12 (14)

Wymiary (ang. dimensions) ustalają kontekst analizy. Przykładowo, analiza sprzedaży czekolady w Auchan, w poszczególnych miesiącach roku jest dokonywana w wymiarze *Produktu*, *Sklepu* i *Czasu*. Inne typowe wymiary to *Lokalizacja* lub *Klient*. Wymiary składają się z **poziomów**, które tworzą hierarchie. Jako przykład można podać wymiar *Lokalizacja* złożony z trzech następujących poziomów: *Sklepy*, *Miasta* i *Województwa*. Wymiar ten ustala hierarchię, w której sklepy należą do miast, a miasta do województw.



Implementacje wielowymiarowego MD

- ROLAP - implementacja w serwerach relacyjnych
 - fakty przechowywane w tabelach faktów
 - wymiary przechowywane w tabelach wymiarów
- MOLAP - implementacja w serwerach wielowymiarowych
 - dane przechowywane w wielowymiarowych tabelach (ang. data cubes), zwanych potocznie kostkami
- HOLAP - implementacja hybrydowa (relacyjno-wielowymiarowa)
 - dane elementarne przechowywane w tabelach
 - dane zagregowane przechowywane w kostkach

ZSBD – wykład 12 (15)

Model wielowymiarowy może być implementowany albo w serwerach relacyjnych (tzw. ROLAP) albo w serwerach wielowymiarowych (tzw. MOLAP). W pierwszym przypadku, dane są składowane w tabelach relacyjnych. W drugim przypadku, są one składowane w wielowymiarowych tablicach. Często w tej samej bazie danych reprezentuje się informacje częściowo w implementacji ROLAP, a częściowo w MOLAP. Taki sposób reprezentacji nazywa się hybrydowym – HOLAP (ang. Hybrid OLAP).



Implementacja ROLAP

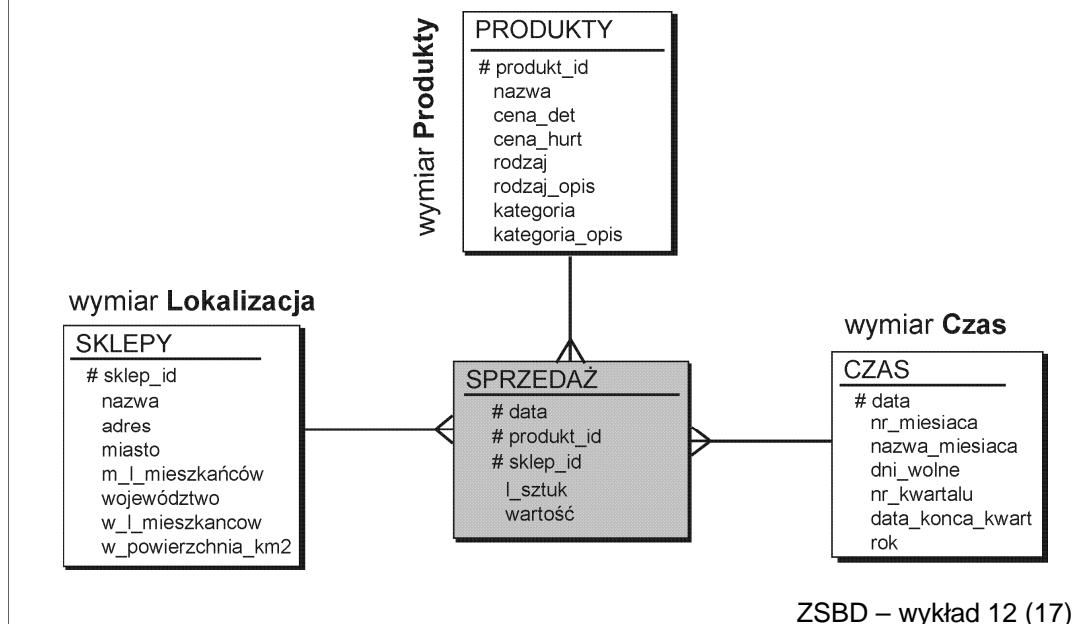
- Schematy podstawowe
 - gwiazda (ang. star schema)
 - płatek śniegu (ang. snowflake schema)
- Schematy pochodne
 - konstelacja faktów (ang. fact constellation schema)
 - gwiazda-płatek śniegu (ang. starflake schema)

ZSBD – wykład 12 (16)

Magazyn danych w technologii ROLAP jest implementowany w postaci tabel, których schemat posiada najczęściej strukturę *gwiazdy* (ang. star schema) lub *płatka śniegu* (ang. snowflake schema) lub *konstelacji faktów* (ang. fact constellation schema) lub strukturę *gwiazda–płatek śniegu* (ang. starflake schema).



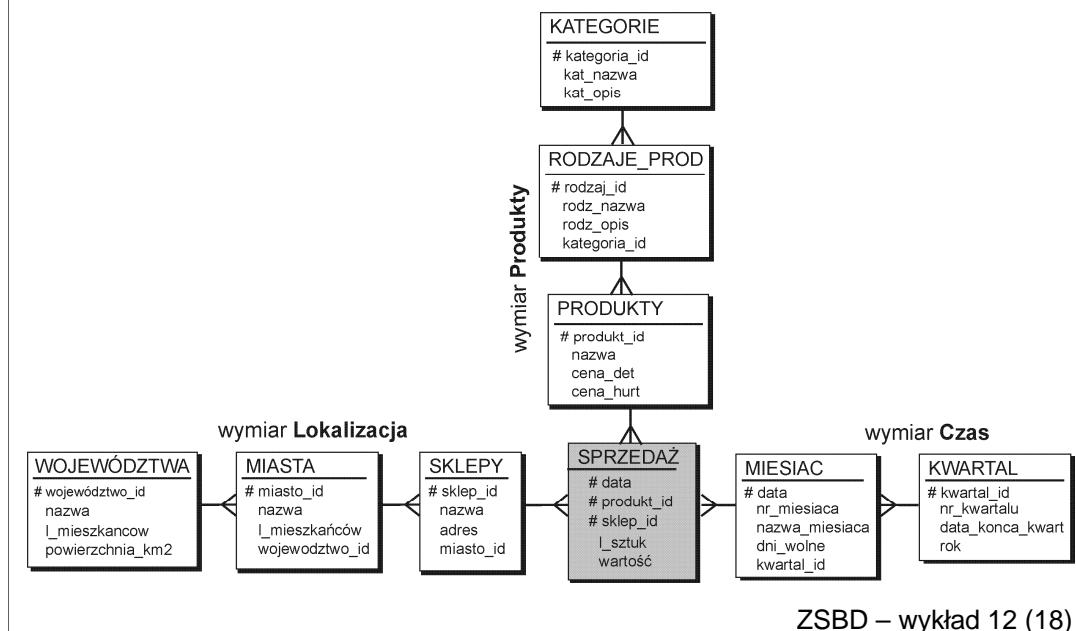
Schemat gwiazdy



Przykładowy schemat *gwiazdy* przedstawia slajd. Centralna tabela *Sprzedaż* zawiera informacje o sprzedaży pewnych produktów, w pewnych sklepach, w określonym czasie. Tabele *Sklepy*, *Produkty* i *Czas* są nazywane tabelami *wymiarów* (ang. dimension tables), natomiast tabela centralna jest nazywana tabelą *faktów* (ang. fact table). Atrybuty miar tabeli *Sprzedaż* to *wartość* i *l_sztuk*. Tabela faktów zawiera również atrybuty *produkt_id*, *sklep_id*, *data*. Są to klucze obce, których wartości wskazują na odpowiednie wymiary. W takim schemacie tabele wymiarów, tj. *Sklepy*, *Produkty* i *Czas* są zdenormalizowane (nie spełniają przynajmniej 3 postaci normalnej).



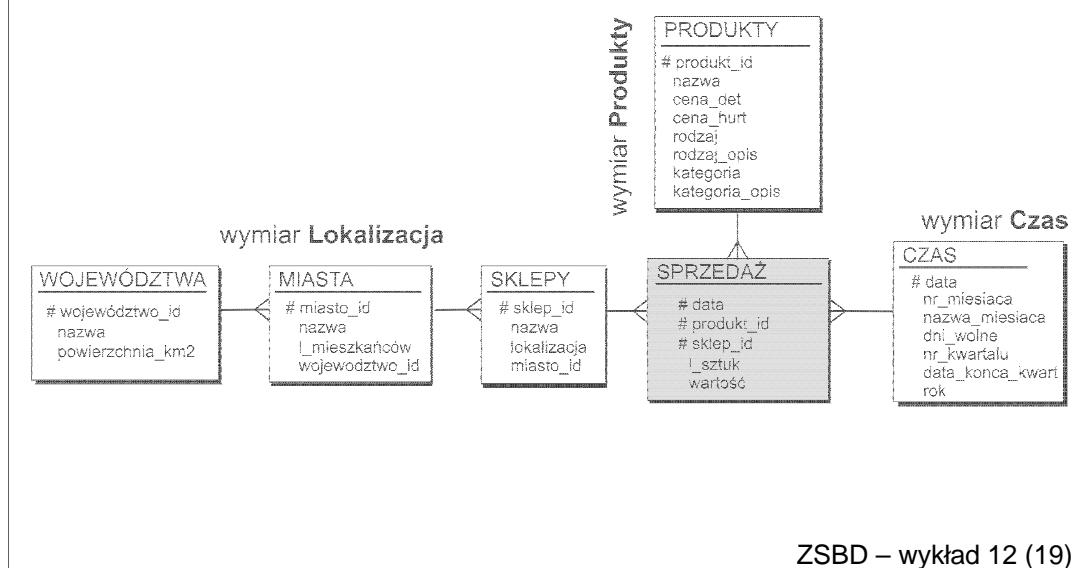
Schemat płatka śniegu



Jeśli wymiary są znormalizowane (spełniają przynajmniej 3 postać normalną), wówczas schemat magazynu danych ma postać *platka śniegu*. Przykładowy schemat o takiej strukturze został przedstawiony na slajdzie. W tym przypadku, wymiary *Lokalizacja*, *Produkty* i *Czas* mają postać hierarchii. Przykładowo, wymiarze *Lokalizacja* każdy sklep (tabela *Sklepy*) znajduje się w mieście (tabela *Miasta*), które z kolei znajduje się w województwie (tabela *Województwa*). Podobnie, w wymiarze produktów istnieje jawną hierarchię, w której produkty należą do rodzajów, a rodzaje do kategorii.



Schemat gwiazda-płatek śniegu



ZSBD – wykład 12 (19)

Natomiast schemat, w którym część wymiarów ma postać znormalizowaną, a część ma postać zdenormalizowaną nazywa się schematem *gwiazdy-płatka śniegu*.

Na slajdzie przedstawiono taki przykładowy schemat. W schemacie tym wymiar *Lokalizacja* jest znormalizowany, a pozostałe dwa wymiary są zdenormalizowane.