1. **Hurtownie danych**
2.

HD definiowana jest na różne sposoby:

-HD to problemowo zorientowany, zintegrowany i stabilny zbiór danych historycznych, który wykorzystywany jest w procesach wspomagania decyzji." (W. H. Inmon)

-HD to baza danych budowana dla potrzeb wspomagania decyzji i utrzymywana oddzielnie od operacyjnych (transakcyjnych) baz danych.

- „Hurtownia danych jest pojedynczą, kompletną i spójną bazą danych utworzoną z różnorodnych źródeł i udostępnioną użytkownikom w sposób, który jest dla nich zrozumiały i użyteczny w kontekście zastosowań biznesowych."
-- Barry Devlin, IBM Consultant

2. Magazynowanie danych
Proces tworzenia i użytkowania hurtowni danych

3. Cechy HD

zorientowana - Zorganizowana wokół głównego przedmiotu zainteresowań, np.: klientów, produktów, sprzedaży; na modelowaniu i analizie danych potrzebnych w procesie wspomagania decyzji a nie na codziennych operacjach

- **zintegrowana** - Tworzona przez integrację wielu, często heterogenicznych, źródeł danych; Stosowane są techniki czyszczenia i transformacji danych; Zapewnienie ujednoliconej konwencji

- stabilna - Stabilność HD oznacza, że jedyna zmiana w jej zawartości to dołączanie nowych danych – stare dane pozostają niezmienione. HD fizycznie odseparowana jest od danych operacyjnych.
Nie występuje operacyjna aktualizacja danych

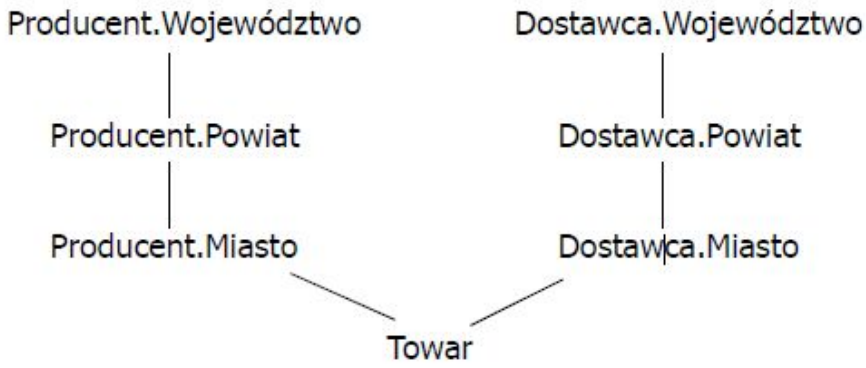
- dane historyczne - Bardzo ważnym atrybutem danych w HD jest czas – możemy mówić o historii danych, a więc ich stanach w różnych momentach czasu

4. Różnica w stosunku do BD

Standardowa BD • Aktualizacja + czytanie • Duża liczba krótkich transakcji • Mb - Gb danych • Bieżący obraz świata • Indeksy wg kluczy • Dane „surowe” • Tysiące użytkowników (np., księgowości, kodr, płac, magazynów, sprzedaży)	HD • Czytanie • Zapytania są długie i złożone • Gb - Tb danych • Historia stanów/operacji • Wiele przeglądań sekw. • Dane wyczyszczone i zagregowane • Setki użytkowników (np., decydenci, analitycy)

5. Wymiary i hierarchie HD

Wymiar: **Towar**



Dwie hierarchie w wymiarze **Towar**:

Towar.Producent:

Towar – Producent.Miasto – Producent.Powiat – Producent.Województwo

Towar.Dostawca:

Towar – Dostawca.Miasto – Dostawca.Powiat – Dostawca.Województwo

6. Schemat gwiazdy

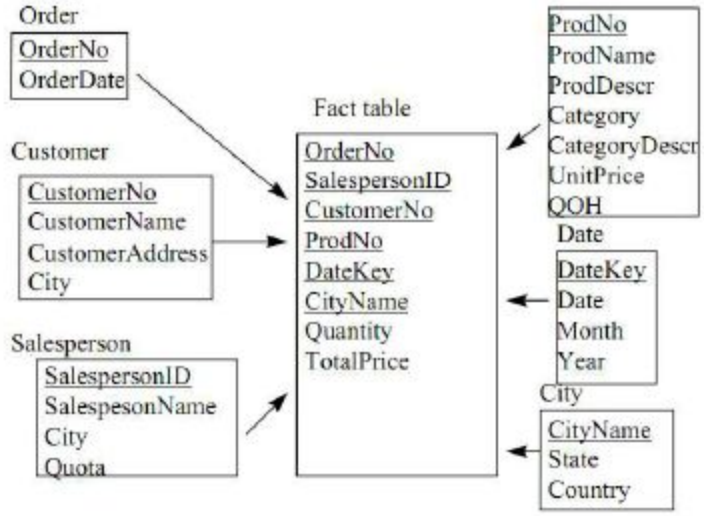
Schemat gwiazdy (ang. star schema) jest najczęściej stosowaną metodą organizacji danych w hurtowni danych. W jego skład wchodzą:

1. Tabeli faktów (ang. fact table): bardzo duży zestaw faktów takich na przykład jak sprzedaż, najczęściej tylko do dołączania.

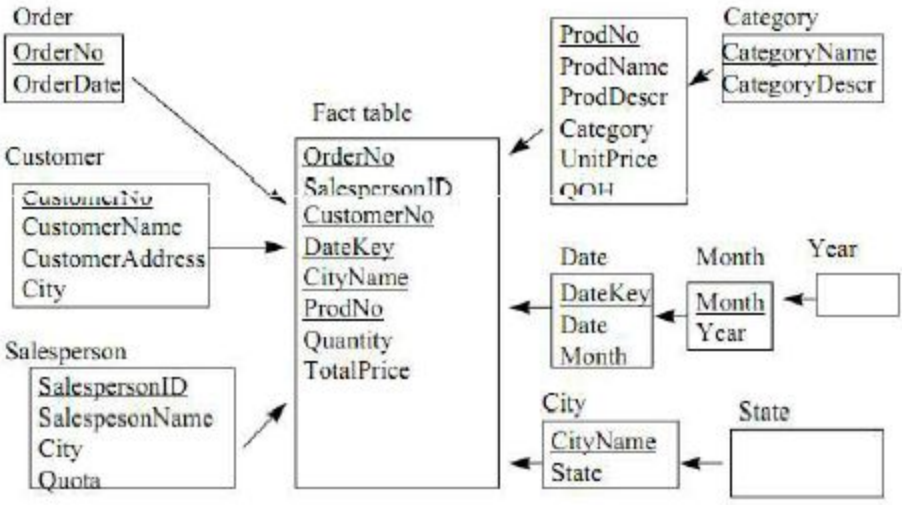
2. Tabele wymiarów (ang. dimension tables): nieduże i najczęściej statyczne tabele zawierające informacje o jednostkach związanych z faktami, na przykład z faktem sprzedaży związane są jednostki: klient, miejsce, towar, moment czasu.

6.1 Schemat gwiazdy i płatka śniegu

gwiazda



płatek



7. Typowe zapytanie OLAP

Zapytania OLAP najczęściej obejmują „star join”, tzn. naturalne złączenie tabeli faktów z wszystkimi lub z większością tabel wymiarów. Może zawierać prosty warunek selekcji.

• Przykład:
SELECT *
FROM Zapłata,Kasa,Karta,Dzień
WHERE
Kasa.Miasto = ‘Poznań’ AND

Zapłata.IdKasy = Kasa.IdKasy AND
Zapłata.IdKarty = Karta.IdKarty AND
Zapłata.Dzień = Dzień.Dzień

8. Implementacje OLAP

Relational OLAP (ROLAP)
• Używany jest relacyjny SZBD do pamiętania danych OLAP i operowania na nich
• Optymalizacja oparta na relacyjnym SZBD
• Największa skalowalność
Multidimensional OLAP (MOLAP)
• Silnik zarządzający pamiętaniem wielowymiarowych tablic
• Szybkie indeksy do obliczania zagregowanych (streszczonych) danych
Hybrid OLAP (HOLAP) (e.g., Microsoft SQL Server)
• Elastyczność: mieszana strategia
Specialized SQL servers (e.g., Redbricks)
• Specjalizowane wspomaganie zapytań SQL na schematach
gwiazda/płatek śniegu star/snowflake schemas

9. Kostka danych
Kostką danych nazywamy zbiór komórek wyznaczony dla wszystkich miar i dla wszystkich możliwych punktów przestrzeni wielowymiarowej zdefiniowanej w schemacie kostki.

- tabela faktów - nazywamy tabelę pamiętaną w bazie danych (hurtowni danych) i zawierającą informacje, które mają podlegać analizie w systemie OLAP. W naszym przypadku jest to tabela

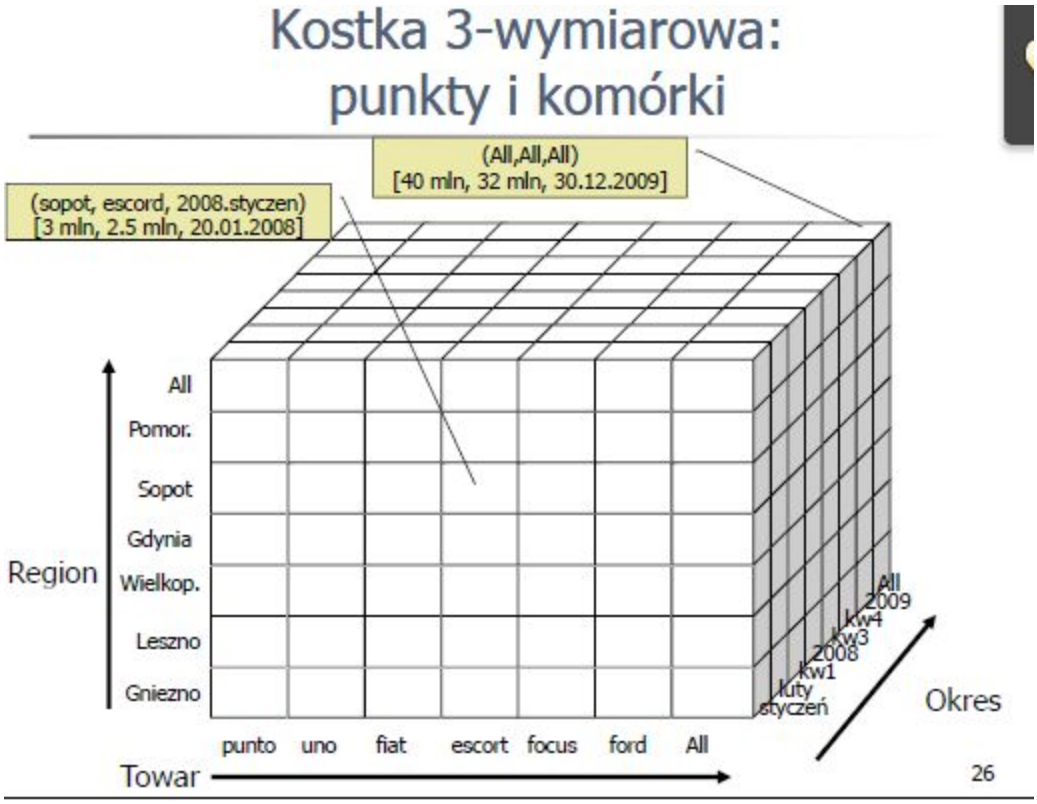
Sprzedaz(IdTow, IdKli, Data, Wartosc, Koszt)

- miary - definiowane są jako te cechy, których wartości są istotne dla użytkownika w procesie analizy. Wyznaczane są na podstawie tabeli faktów. Intuicyjnie rzecz biorąc, wartości te uzyskiwane są w wyniku „pomiarów” w przestrzeni wielowymiarowej.
Np.
Sprzedaz(IdTow, IdKli, Data, Wartosc, Koszt)
Przychod : Sum(Wartosc),
Koszt : Sum(Koszt),
Zysk : Przychod – Koszt,
DataOstSprz : Max(Data),
LiczbaTrans : Count(IdKli),
LiczbaKli : CountDistinct(IdKli).

- wymiary - definiują przestrzeń, w której analizowane są miary.

Każdy wymiar organizowany jest hierarchicznie poprzez określenie:
• hierarchii (opcjonalnie) (ang. hierarchy),
• poziomów (ang. levels)
• członów, elementów (ang. members).

10. Schemat 3D kostki



11. MDX - teoria

- Multidimensional Expressions (MDX) jest językiem zapytań używanym do przetwarzania danych wielowymiarowych w Microsoft SQL Server 2000/2005/2008/2012 Analysis Services (SSAS).

- MDX oparty jest na specyfikacji XMLA (XML for Analysis) rozszerzoną na specyfikę serwera SQL Server 2005 Analysis Services

Zapytania i wyrażenia języka MDX używane są do:
- udostępnienia aplikacjom klienta danych zawartych w kostce;
- formatowania wyniku;

- wykonania zadań projektowych na kostce
- Wykonania zadań administracyjnych takich jak określenie bezpieczeństwa dla wymiarów i komórek.

12. MDX zapytanie SELECT

Wyrażenie MDX SELECT ma następującą składnię:

[WITH <specyfikacja_formuły>[, <specyfikacja_formuły> ...]]
SELECT [<specyfikacja_osi> [, <specyfikacja_osi>...]]
FROM [<specyfikacja_kostki>]
[WHERE [<specyfikacja_plastra>]]

Składnia:

SELECT [<specyfikacja_osi> [, <specyfikacja_osi>...]]
<specyfikacja_osi> ::= <zbiór_ciągów_członów> ON <nazwa_osi>
<nazwa_osi> ::= COLUMNS | ROWS | PAGES | SECTIONS | CHAPTERS| AXIS(<indeks>)

FROM [<specyfikacja_kostki>]
<specyfikacja_kostki> ::= <nazwa_kostki>

[WHERE [<specyfikacja_plastra>]]
<specyfikacja_plastra> ::= <punkt>

np.
WHERE ([Okres].[OkresSprz].[Rok].&[2005],[Measures].[LiczbaKlientow])

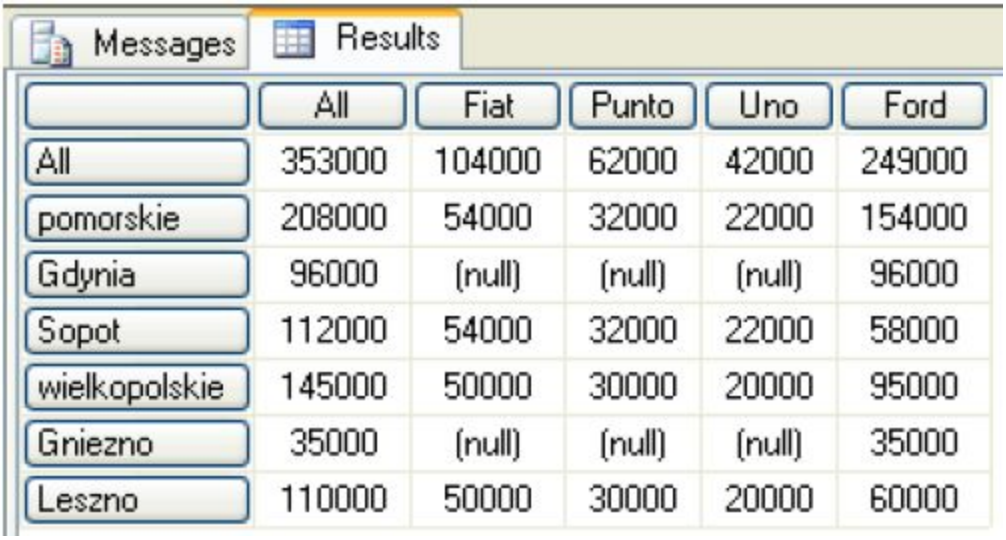
We frazie WITH można definiować miary obliczane oraz przypisywać nazwy zbiorom. Składnia tej frazy jest następująca:
WITH <specyfikacja_formuły> [, <specyfikacja_formuły> ...]
<specyfikacja_formuły> ::= <specyfikacja_członu> | <specyfikacja_zbioru>
<specyfikacja_członu> ::= MEMBER <rodzic_członu>.<nazwa_członu> AS '<wyrażenie>'
<specyfikacja_zbioru> ::= SET <nazwa_zbioru> AS '<zbiór>'

np.
WITH MEMBER Measures.ZyskProc AS '(Przychod - Koszt)/Koszt * 100'

13. MDX - przykład zapytania

Napisz w języku MDX polecenie tworzenia tabeli przestawnej (pivot table), w której zawarte są informacje o wartości sprzedaży samochodów (patrz poniżej) we wszystkich możliwych regionach.

Tabela wynikowa ma mieć następującą postać:



Messages | Results

	All	Fiat	Punto	Uno	Ford
All	353000	104000	62000	42000	249000
pomorskie	208000	54000	32000	22000	154000
Gdynia	96000	(null)	(null)	(null)	96000
Sopot	112000	54000	32000	22000	58000
wielkopolskie	145000	50000	30000	20000	95000
Gniezno	35000	(null)	(null)	(null)	35000
Leszno	110000	50000	30000	20000	60000

Przykład:
Osi kolumn przypisujemy człony hierarchii SamochodHier wymiaru SamochodDim
Osi wierszy przypisujemy człony hierarchii RegionHier wymiaru RegionDim
Domyślnie przyjmuje się pierwszą miarę - Przychod
W komórkach podane są wartości miary Przychod odpowiadające punktom (SamochodHier, RegionHier) – domyślnym członem wymiaru Okres jest Okres.OkresSprz.[All].

SELECT [<specyfikacja_osi> [, <specyfikacja_osi>...]]
FROM [<specyfikacja_kostki>]
[WHERE [<specyfikacja_plastra>]]

SELECT {[SamochodDim].[SamochodHier].[All],Fiat,Punto,Uno,Ford} ON COLUMNS,
{[RegionDim].[RegionHier].[All],pomorskie, Gdynia, Sopot,
wielkopolskie, Gniezno, Leszno} ON ROWS
FROM SalonSamCube

15. Metadane w HD
Metadane określają jakie dane i gdzie się znajdują
Metadane mogą zawierać następujące składniki:

- słowniki danych – definicje obsługiwanych baz danych i relacji między elementami danych,
- przepływy danych – kierunek i częstotliwość przekazywania danych w systemie

- transformacje danych – operacje na danych podczas ich przenoszenia
- wersje danych – numery wersji danych i informacje o ich modyfikacjach,
- profile danych – statystyki użycia danych
- nazwy danych – nazwy nadane poszczególnym polom danych
- uprawnienia użytkowników – dotyczące dostępu do danych

16. Rodzaje metadanych

W repozytorium metadanych znajdują się:

- metadane z perspektywy pojęciowej (dane biznesowe),
- metadane z perspektywy logicznej (schemat),
- metadane z perspektywy fizycznej,
- statystyki danych,
- statystyki użycia,
- informacje administracyjne.

17. HD - OLTP a OLAP

W tradycyjnym ujęciu – HD to wielkie magazyny danych historycznych znajdujący się pomiędzy

- danymi źródłowymi w systemach OLTP
- systemami wspomagania decyzji i eksploracji danych OLAP

Przemawiają za tym trzy argumenty – nie mieszać OLTP i OLAP(!)

<table>
<tr><td colspan="2">różnice w aspektach wydajność i dostępności:</td></tr>
<tr><td>- OLTP</td><td>- OLAP</td></tr>
<tr><td>- mają znaczenie krytyczne dla działania firmy;</td><td>- dostęp do ogromnych zbiorów danych – można stosować zrównoleglanie</td></tr>
<tr><td colspan="2">różnice w modelowaniu pojęciowym:</td></tr>
<tr><td>- korzystanie z modelu relacyjnego, czasem semistrukturalnego,</td><td>- myślenia w pojęciach arkuszy kalkulacyjnych, tj. wielowymiarowych danych o bogatej strukturze</td></tr>
<tr><td colspan="2">niezgodność czasu i poziomu ogólności:</td></tr>
<tr><td>- dane aktualne, operacyjne, na poziomie szczegółowym</td><td>- dane historyczne na wyższym poziomie ogólności</td></tr>
</table>

różnice w aspektach wydajność i dostępności:

- OLTP mają znaczenie krytyczne dla działania firmy;
- OLAP dostęp do ogromnych zbiorów danych – można stosować zrównoleglanie

różnice w modelowaniu pojęciowym:

- OLTP – korzystanie z modelu relacyjnego, czasem semistrukturalnego,
- OLAP – myślenia w pojęciach arkuszy kalkulacyjnych, tj. wielowymiarowych danych o bogatej strukturze

niezgodność czasu i poziomu ogólności:

- OLTP – dane aktualne, operacyjne, na poziomie szczegółowym
- OLAP – dane historyczne na wyższym poziomie ogólności