Hurtownie danych - przegląd technologii

Robert Wrembel
Politechnika Poznańska
Instytut Informatyki
`Robert.Wrembel@cs.put.poznan.pl`
`www.cs.put.poznan.pl/rwrembel`

Kierunki rozwoju

- Półautomatyczne konstruowanie schematów HD
- Ewolucja HD ⇨ temporalne i wielowersyjne HD
- ETL
  - optymalizacja ETL
  - ETL czasu rzeczywistego
  - ewolucja ETL
  - konstruowanie ETL dla źródeł o złożonych strukturach
- Magazynowanie i przetwarzanie danych złożonych
  - HD XML
  - przestrzenny OLAP (Spatial OLAP)
  - analiza danych strumieniowych (data streams)
  - HD dla bio-informatyki
- Integracja HD

Konstruowanie schematu HD

- **Modelowanie konceptualne**
  - **model związków-encji**
  - **UML i rozszerzenia stereotypów**
- **Koncentracja na wymagania użytkowników (user/demand driven) - podejście tradycyjne**
  - **analiza wymagań**
  - **wywiady z użytkownikami realizowane przez analityków**
  - **uwzględnia cele biznesowe**
- **Koncentracja na strukturę i zawartość źródeł (source/supply/data driven)**
  - **schemat HD odzwierciedla strukturę źródeł**
  - **automatyczne konstruowanie**
  - **nie uwzględnia celów biznesowych**

Konstruowanie schematu HD

- **Półautomatycznie na podstawie struktury systemów źródłowych**

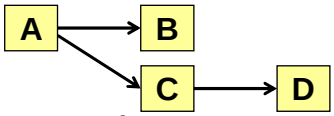
- Song I.-Y., Khare R., Dai B.: SAMSTAR: a semi-automated lexical method for generating star schemas from an entity-relationship diagram. DOLAP, 2007
- Romero O., Abello A.: Automating Multidimensional Design from Ontologies. DOLAP, 2007
- Phipps D., Davis K.: Automating Data Warehouse Conceptual Schema Design and Evaluation. DMDW, 2002
- Jensen M, Holmgren R., Pedersen T.B.: Discovering Multidimensional Structure in Relational Data. DAWAK, 2004

Konstruowanie schematu HD

- SAMSTAR
  - półautomatyczne generowanie schematu ER HD
  - analiza struktury schematu źródłowego (zw-encji) + korekta projektanta
- Kroki
  - znalezienie faktów i bezpośrednich poziomów (automat.)
  - znalezienie przechodnich poziomów (automat.)
  - wybór faktów (projektant)
  - wybór wymiarów dla wskazanych faktów (automat.)
    - wykorzystanie WordNet - poszukiwanie synonimów
    - wykorzystanie Dimensional Design Pattern (DDP) - określenie struktury wymiarów
    - DDP - zbiór ponad 100 wzorców typowych wymiarów

SAMSTAR

- Reguły wyboru faktów i poziomów
  - encje po stronie związków kardynalności M ⇨ kandydaci faktów
  - encje po stronie związków kardynalności 1 ⇨ kandydaci poziomów
- Bezpośredni i pośredni poziom
- CTV (Connection Topology Value)
  - suma ważona poziomów bezpośrednich i pośrednich (tranzytywnie dla pośrednich)
  - waga poziomu bezpośredniego > waga poziomu pośredniego
  - obliczana dla każdej encji w schemacie źródłowym

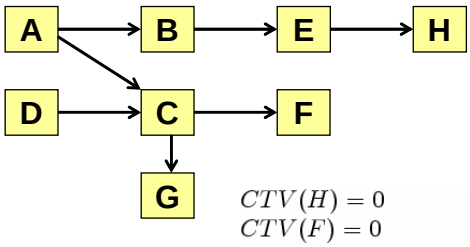


$$CTV(e) = 1 \cdot w_{direct} + w_{indirect} \cdot \sum_{i=1}^{n} CTV(i)$$

SAMSTAR

- **Obliczenie CTV**

wagi: $w_{direct} = 1$; $w_{indirect} = 0.8$



$$CTV(H) = 0$$
$$CTV(F) = 0$$
$$CTV(G) = 0$$
$$CTV(E) = 1 * 1 + 0.8 * CTV(H) = 1$$
$$CTV(B) = 1 * 1 + 0.8 * CTV(E) = 1.8$$
$$CTV(C) = 1 * 2 + 0.8(CTV(G) + CTV(F)) = 2$$
$$CTV(D) = 1 * 1 + 0.8(CTV(C)) = 2.6$$
$$CTV(A) = 1 * 2 + 0.8 * (CVT(B) + CTV(C)) = 5.04$$

SAMSTAR

- **Encje dla których CTV > ϑ są encjami faktów**
- **Pozostałe encje to kandydaci do wymiarów**

$$\vartheta = \frac{\sum_{i=1}^{n} CTV(i)}{n} + k \cdot StDev$$

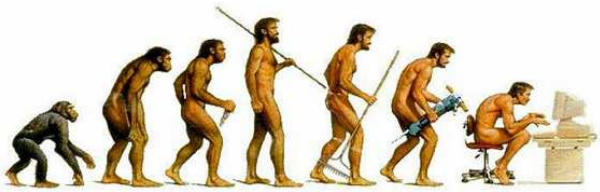
  - **k współczynnik systemowy, dobierany przez projektanta; wraz ze wzrostem wartości maleje liczba proponowanych encji faktów (w praktyce 1.5 - 1.75)**
- **Pozostałe encje to kandydaci do wymiarów**
- **Encje faktów mogą mieć tę samą semantykę ale różne nazwy ⇨ zastosowanie WordNet do poszukiwania synonimów**

SAMSTAR

- Kandydaci wymiarów to:
  - encje poziomów bezpośrednich od strony związku 1 powiązane związkiem kardynalności M z encjami faktów
  - encje poziomów pośrednich
- Budowanie wymiarów
  - encje poziomów mogą mieć tę samą semantykę ale różne nazwy ⇨ zastosowanie WordNet do poszukiwania synonimów
  - zastosowanie Dimension Design Pattern do zbudowania hierarchii wymiarów
  - dodanie wymiaru czasu na podstawie DDP
- Akceptacja schematu przez projektanta (manualnie)
  - ewentualna modyfikacja otrzymanego schematu przez projektanta

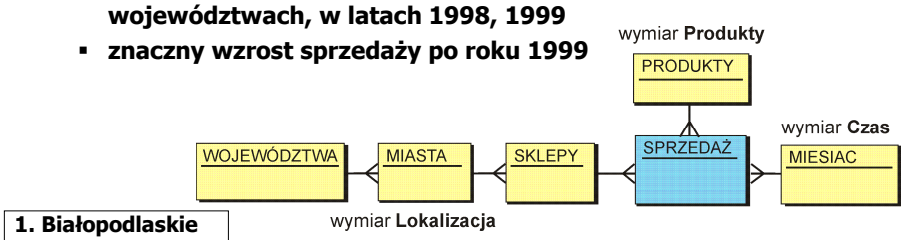
Ewolucja HD

- Zmiany w strukturze/schemacie źródła danych
- Zmiany organizacyjne (podział terytorialny, struktura organizacyjna, reklasyfikacja)
- Zmiany wymagań użytkowników
- Symulowanie scenariuszy biznesowych



Ewolucja HD - przykład (1)

- **Zmiana granic administracyjnych województw**
  - **porównanie sumy sprzedaży czekolady w poszczególnych województwach, w latach 1998, 1999**
  - **znaczny wzrost sprzedaży po roku 1999**



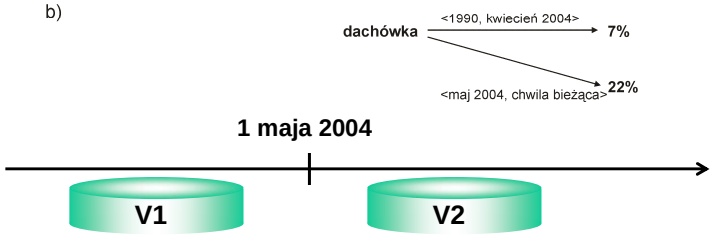
1. Białopodlaskie	**1. Dolnośląskie**
2. Białostockie	**2. Kujawsko-Pomorskie**
3. Bielskie	**3. Lubelskie**
...	**...**
...	**14. Warmińsko-Mazurskie**
...	**15. Wielkopolskie**
47. Wrocławskie	**16. Zachodniopomorskie**
48. Zamojskie	
49. Zielonogórskie	

Ewolucja HD - przykład (2)

- **Przyporządkowanie produktów do kategorii**

a)

b)



Ewolucja HD

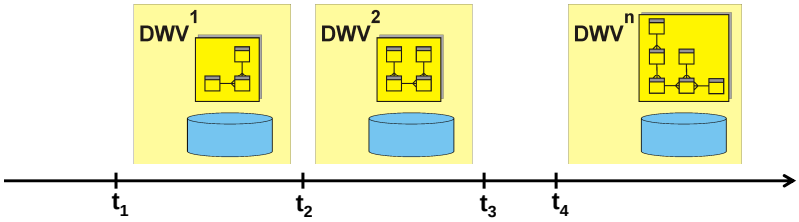
- Przewidywanie przyszłości i trendów biznesowych
- Analiza alternatywnych rozwiązań biznesowych (ang. what-if analysis)
- Przykład:
  - zapytanie o spadek/wzrost łącznej kwoty mandatów płaconych w województwie wielkopolskim, przy założeniu, że minimalna i maksymalna grzywna za jazdę bez zapiętych pasów bezpieczeństwa została zwiększona o 10%

Temporalne HD

- Znaczniki czasowe ważności danych
- Umożliwiają przechowywanie historycznych wersji DANYCH
- Wersje uporządkowane liniowo
  - brak wsparcia dla symulacji

Wielowersyjna HD

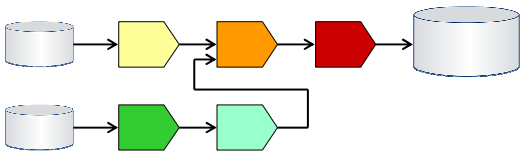
- Mechanizm rozwiązującego problemy związane z koniecznością zarządzania zmianami schematu i struktury wymiarów
- MVDW składa się ze zbioru trwałych wersji
  - każda wersja posiada znaczniki czasowe początku i końca jej ważności



Temporalne i wielowersyjne języki zapytań

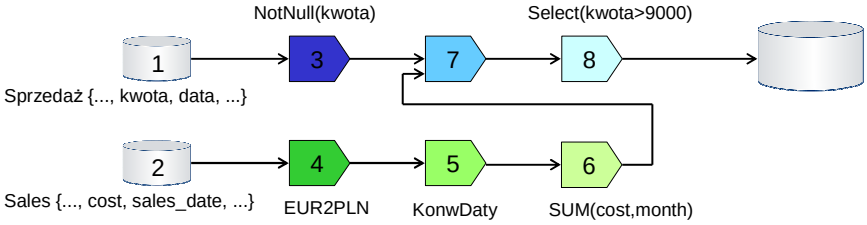
- Przeszukiwanie danych w wersjach w zadanym przedziale czasowym
- Wyznaczenie wyników zapytań z poszczególnych wersji i ich integracja w jeden spójny zbiór posiadający strukturę magazynu z zadanej wersji lub momentu czasowego

Optymalizacja ETL



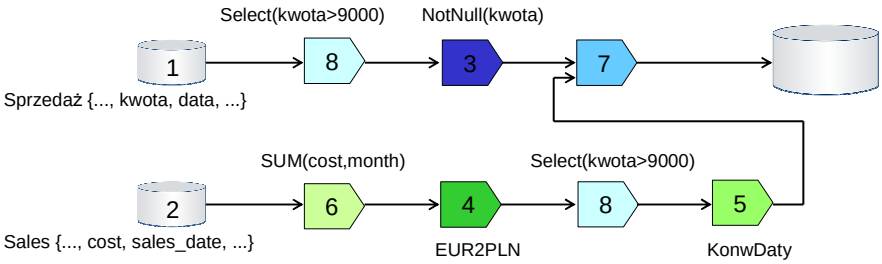
- **Optymalizacja przez transformację przepływu**
  - **zmianę kolejności elementów w przepływie**
  - **zrównoleglenie zadania**
  - **scalenie kilku zadań**
- **Wyznaczenie poprawnych transformacji dla zadanego przepływu**
- **Znalezienie przepływu minimalizującego czas wykonania**

Przykład



- **Źródło Sprzedaż**
  - **kwota [PLN]**
  - **data [yyyy-mm-dd]**
  - **przechowuje dane nt sprzedaży miesięcznej**
- **Źródło Sales**
  - **cost [EUR]**
  - **sales_date [dd/mm/yy]**
  - **przechowuje dane nt sprzedaży dziennej**

Przykład



- **Przepływ niebieski ⇨ selekcja jak najwcześniej (przed 3)**
- **Przepływ zielony**
  - **selekcja dopiero po konwersji waluty i wyliczeniu sumy sprzedaży miesięcznej**
  - **wyliczenie SUM(cost, month) możliwe przed EUR2PLN**
  - **konwersja daty po odfiltrowaniu rekordów**

Optymalizacja ETL

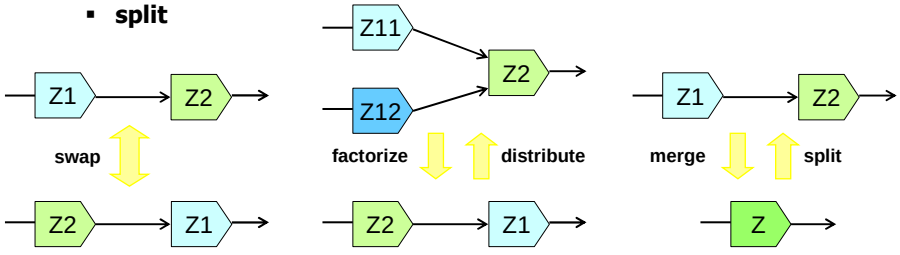
- **Problem**
  - **transformacje wyrażone w algebrze relacji (selekcja, projekcja, połączenie, op. na zbiorach) ⇨ optymalizacja algebraiczna**
  - **trudna optymalizacja funkcji**
    - **mogą być implementowane w różnych językach programowania**
    - **trudność oszacowania kosztu wykonania**
- **Systemy komercyjne**
  - **przepływy ETL definiowane przez projektanta (odpowiedzialny za optymalizację)**
  - **optymalizacja tylko poleceń SQL w ramach zadań ETL**
    - **optymalizacja w ramach pojedynczego zadania**

Optymalizacja ETL

- Wykonywanie niektórych operacji w systemie operacyjnym (poza bazą danych)
  - wywołanie zewnętrznych funkcji transformacji
- Zmiana kolejności wykonywania operacji
- Minimalizacja rekordów do przetwarzania
  - filtrowanie jak najwcześniej
- Wykorzystanie (materializowanie) wyników pośrednich
- Przetwarzanie równoległe
- Wykorzystanie metadanych opisujących źródło
  - `select * from klienci@ZR1 where miasto='Poznań'`
  - czy jest indeks na atrybucie miasto?
  - jaka jest selektywność warunku
  - jaki optymalizator wykorzystuje źródło
  - odczytanie całej tabeli klienci może okazać się bardziej efektywne

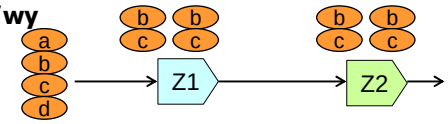
Arktos II

- Transformacje przepływów
  - swap - selekcja jak najwcześniej
  - factorize - Z11 i Z12 wykonują te same operacje na 2 różnych strumieniach wejściowych ⇨ wykonanie operacji na scalonym strumieniu
  - distribute
  - merge - logiczne grupowanie zadań, które muszą nastąpić po sobie
  - split

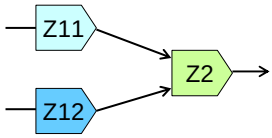


Arktos II

- Poprawność transformacji
- Swap
  - Z1 ma jedno źródło, Z2 ma jeden cel
  - kompatybilność schematów we/wy
    - we.Z1={b,c} i wy.Z1={b,c}
    - we.Z2={b,c} i wy.Z2={b,c}



- Factorize/Distribute
  - Z11 i Z12 mają 1 cel Z2 (operacja na zbiorach)
  - Z11 i Z12 realizują tę samą operację ale na innych przepływach wejściowych



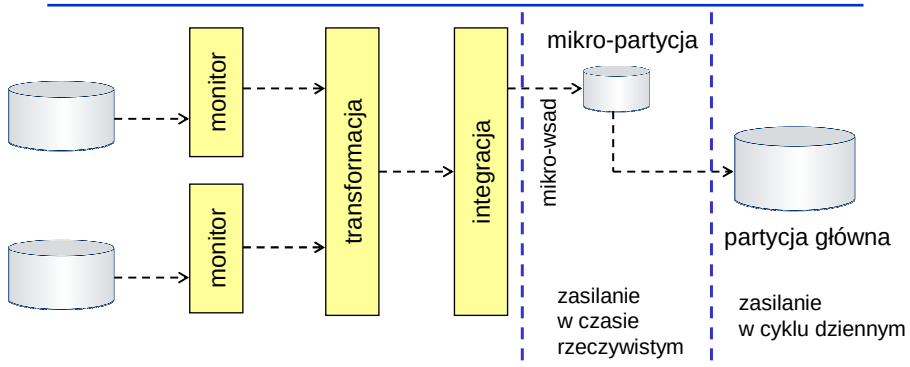
Arktos II

- Przeszukiwanie przestrzeni dozwolonych transformacji przepływu ETL
  - pełne (nie realizowalne w skończonym czasie dla 40 i więcej zadań)
  - heurystyki

ETL czasu rzeczywistego

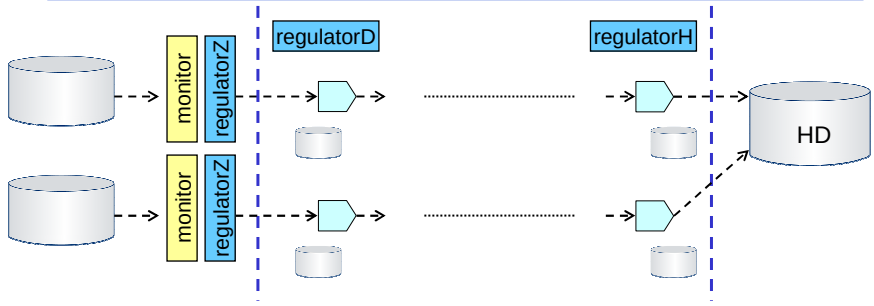
- **Zastosowanie ⇨ HD czasu rzeczywistego**
- **Cechy**
  - **czas pomiędzy zmianą w źródle, a uaktualnieniem HD ⇨ kilka - kilkadziesiąt minut**
  - **wolumen odczytywanych i przetwarzanych danych ⇨ mały w porównaniu z podejściem standardowym ⇨ możliwość przetwarzania potokowego w RAM**
- **Problemy**
  - **inny rodzaj przetwarzania ⇨ mikro-wsadowe**
  - **częstotliwość uaktualniania HD i struktur fizycznych (perspektywy zmaterializowane, indeksy)**

ETL czasu rzeczywistego



- **Mikro-partycja (MP) przechowuje zmiany z bieżącego dnia**
- **Zawartość MP przesyłana wsadowo do partycji głównej (PG) np. raz na dobę**
- **Pełen obraz danych ⇨ MP+PG (zintegrowane np. za pomocą perspektywy)**
- Kimball R., Caserta J.: The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning. Wiley, 2004

ETL czasu rzeczywistego



- **regulatorZ - zarządza przesyłaniem danych ze źródła, (bada aktualne obciążenie źródła, czy aktualny wolumen danych zmieści się w zadanym oknie czasowym)**
- **regulatorD - informuje, z którego źródła dane są gotowe do odczytania (przygotowany został wolumen danych)**
- **regulatorH - zarządza przesyłaniem danych do HD (zapewnia QoS-QoD)**
- Vassiliadis P., Simitsis A.: Near Real Time ETL. Annals of Information Systems, Springer, 2009

Ewolucja ETL

- **Zmiana struktury źródła danych**
  - **konieczność przedefiniowania (fragmentu) procesu ETL**
  - **ponowna optymalizacja procesu**
- **Problematyka**
  - **wykrywania zmian w strukturze źródeł**
  - **automatyczna modyfikacja procesu ETL**
- Sellis T., Simitsis A.: ETL Workflows: From Formal Specification to Optimization. ADBIS 2007

ETL dla danych złożonych

- Źródła danych
  - multimedialne, GIS, XML, tekstowe bd, strony WWW
- Złożona struktura danych
  - dźwięk, obraz, mapa, dokument XML, dokument tekstowy
- Czyszczenie danych i eliminowanie duplikatów
  - równoważne sobie obrazy, mapy
  - równoważne dokumenty XML (struktura i/lub zawartość)
- Dane o większych rozmiarach
- Przetwarzanie danych bardziej złożone obliczeniowo ⇨ problem zakończenia procesu w zadanym czasie

Magazynowanie i analiza danych złożonych

- Obiektowo-relacyjne HD
- Multimedialne HD
- Semistrukturalne HD
- Magazynowanie i przetwarzanie danych ze strumieni (ang. data streams)

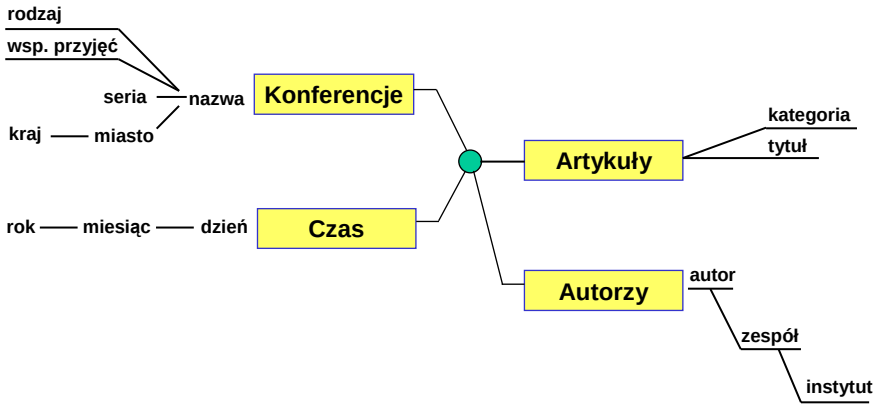
Hurtownie danych dla XML

- Powszechność danych XML (systemy e-...)
- Potrzeba analizy tych danych w sposób podobny do tradycyjnego
- Budowanie HD w oparciu o źródła XML
- Analiza danych XML

- Byung-Kwon Park B-K., Han H., Song I-Y.: XML-OLAP: A Multidimensional Analysis Framework for XML Warehouses. DAWAK, 2005
- Boussaid O., Messaoud R.B., Choquet R., Anthoard S.: X-Warehousing : An XML-Based Approach for Warehousing Complex Data. ADBIS, 2006
- Ravat F., Teste O., Tournier R., Zurlfluh Z.: Designing and Implementing OLAP Systems from XML Documents. Annals of Information Systems, Springer, 2008

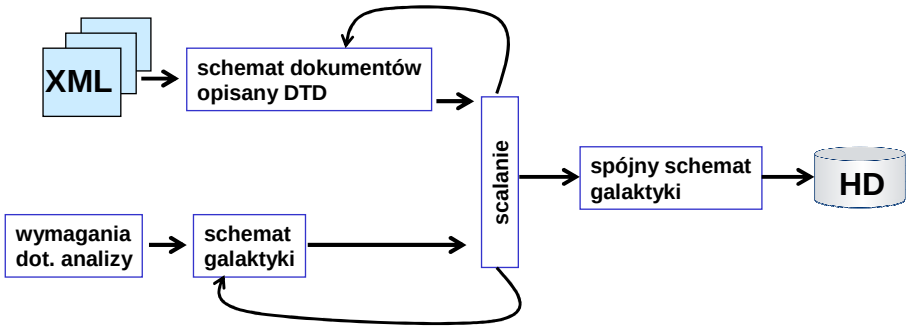
Hurtownie danych dla XML

- Schemat konceptualny HD dla XML jest reprezentowany tzw. modelem galaktyki



Hurtownie danych dla XML

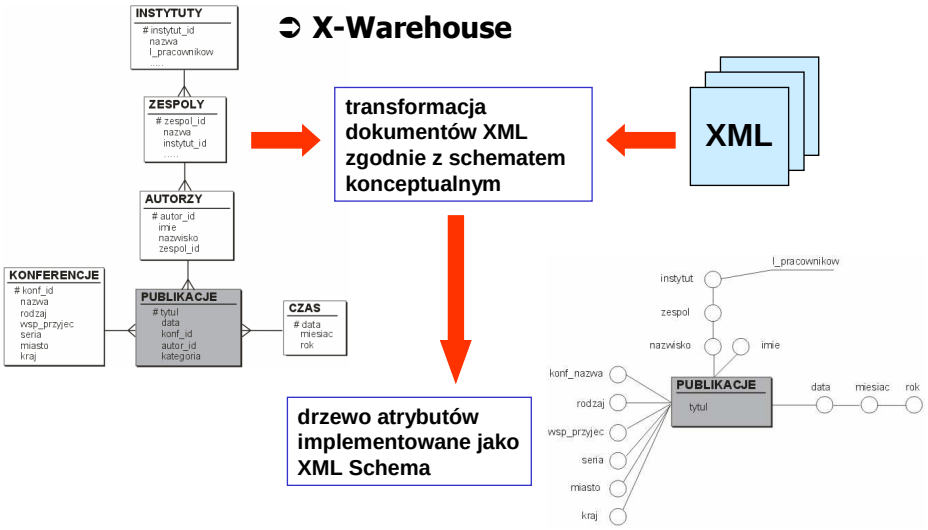
- Metodyka projektowania HD



- Scalanie schematu galaktyki ze schematem dokumentów XML może wymagać modyfikacji galaktyki i dokumentów
- Dokumenty XML składowane w relacyjnej bazie danych

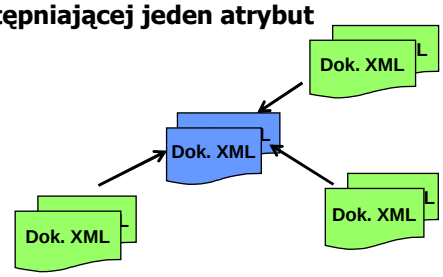
Hurtownie danych dla XML

- X-Warehouse



Hurtownie danych dla XML

- **XML-OLAP**
  - **fakty i instancje wymiarów są reprezentowane za pomocą dokumentów XML**
  - **przechowywane w bazie danych XML**
  - **język analizy danych XML-MDX bazujący na językach MDX i XQuery**
  - **miary specyfikowane za pomocą wyrażeń XQuery - analogia do perspektywy udostępniającej jeden atrybut**
  - **operatory agregacji miary**
    - **dla wartości numerycznych**
    - **dla wartości tekstowych**
      - podsumowanie treści
      - główny temat
      - n słów kluczowych

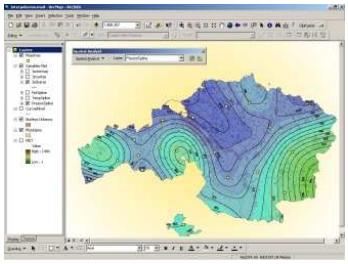


Przestrzenny OLAP

- **80% danych ma charakter przestrzenny** [Gonzales L.: Seeking Spatioal Intelligence. Intelligent Enterprise Magazine, 2(3), 2000]
  - **obiekty geometryczne 2 i 3 wymiarowe**
  - **mapy, zdjęcia**
- **Systemy GIS (Geographical Information Systems)/przestrzenne bazy danych**
  - **wsparcie dla przetwarzania obiektów przestrzennych**
  - **indeksowanie**
  - **składowanie**
  - **brak wparcia dla OLAP**

Przestrzenny OLAP

- Analiza danych przestrzennych
  - geografia, geologia, urbanistyka, gospodarka leśna, przemysł wydobywczy, astronomia
- Analiza przestrzenno-czasowa
  - np. zmiany biegu rzek, ukształtowania terenu
- Przetwarzanie obrazów + analiza ich zawartości



Przestrzenny OLAP

- Zbiory danych przestrzennych
  - United Nations Environment Program
    - zasoby wodne, populacje, obszary leśne, emisja zanieczyszczeń, zmiany klimatyczne
  - NASA
  - System obserwacji ziemi NASA EOSDIS
    - głównie zdjęcia satelitarne
    - 1 000 TB rocznie

National Aeronautics and Space Administration

Przestrzenny OLAP

- Wymiar
  - tradycyjny (dane znakowe i numeryczne)
  - przestrzenny
  - mięszany (poziom najniższy przestrzenny, poziomy wyższe znakowe i numeryczne)
- Miara
  - numeryczna
  - przestrzenna (np. kolekcje obiektów lub wskaźników do obiektów)

Problematyka P-OLAP

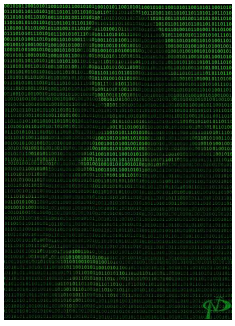
- Integracja danych
  - źródła w różnych formatach: mapy rastrowe i wektorowe, modele obiektowe, relacyjne
- Nowe modele HD
- Nowe operacje analityczne
- Efektywność analiz
  - struktury fizyczne
- Techniki wizualizacji
  - na mapie politycznej, topograficznej
  - zmienny poziom szczegółowości

Analiza danych strumieniowych (1)

- Charakterystyka
  - dane napływające ciągle
  - ogromne ilości
- Źródła
  - sensory (procesy technologiczne, eksperymenty fizyczne)
  - sygnały telekomunikacyjne
  - operacje myszką w aplikacjach internetowych
  - kursy giełdowe

Analiza danych strumieniowych (2)

- Problemy
  - składowanie danych historycznych
    - agregaty
    - reprezentatywne próbki
  - ciągła analiza danych (continuous queries)
    - w oknie czasowym
  - zapytania ad-hoc jak w tradycyjnej bd
  - analiza danych historycznych i bieżących
  - efektywne wyszukiwanie danych
    - indeksowanie



Analiza danych strumieniowych (3)

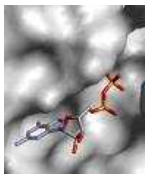
- Problemy cd.
  - Zapewnienie jakości danych (Quality of Service)
    - dane przybliżone (inteligentne budynki)
    - dane dokładne (sensory ognia i dymu)
  - Efektywność przetwarzania
    - minimalizacja zajętości pamięci operacyjnej
    - maksymalizacja przepustowości
    - minimalizacja czasu przetwarzania danych

Analiza danych strumieniowych (4)

- QoS, efektywność przetwarzania zapytań, architektury, mechanizmy aktywne
  - NIAGARA, Fjord, Aurora, STREAM (Stanford Univ.)
- Inteligentny dom (smart home)
  - Intelligent Home Project (Massachussets Univ.)
  - Georgia Tech Aware Home
  - MIT Intelligent Room

MD dla bioinformatyki

- Dane o złożonej strukturze (grafowej)
  - łańcuchy genetyczne
  - struktury białek
  - struktury wiązań chemicznych
- Problem efektywnego składowania
- Analiza struktur złożonych
- Eksploracja struktur złożonych



Integracja HD

- Heterogeniczność i rozproszenie