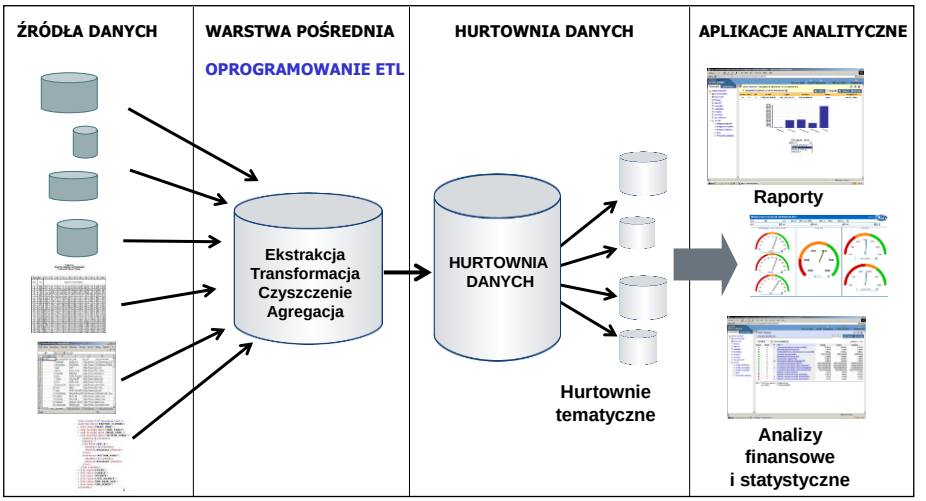
Hurtownie danych - przegląd technologii

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

Zasilanie danymi - procesy ETL

- Charakterystyka źródeł danych
- ETL w architekturze HD
- Charakterystyka ETL
- Ekstrakcja
- Transformacja
- Wczytanie
- Wymagania dla ETL
- Metadane ETL

ETL w architekturze HD



Charakterystyka ETL

- **Konstruowanie procesów ETL**
  - **krytyczne dla działania HD**
    - **jakość danych**
    - **aktualność danych**
    - **zasilanie w ściśle określonym oknie czasowym (opóźnienia skutkują niedostępnością HD)**
  - **kosztowne i czasochłonne**
    - **do 70% zasobów projektowych**
      - **ludzie**
      - **sprzęt**
      - **oprogramowanie**

Charakterystyka ETL

- **Raport Gartnera nt. projektów HD w instytucjach finansowych Fortune 500**
  - **100 osób zaangażowanych w projekt HD**
  - **55 ETL**
  - **17 administratorzy systemu (BD, sprzęt)**
  - **4 architektów systemu**
  - **9 konsultanci dla użytkownika końcowego od strony technologii BI**
  - **5 programistów**
  - **9 menedżerów**
  - **sprzęt**
    - serwery wieloprocesorowe, dyski TB (5 mln USD)
    - oprogramowanie ETL (1 mln USD)
  - **typowa liczba źródeł danych ⇨ 10-50**

Problematyka naukowo-badawcza

- **Przetwarzanie dużych wolumenów danych w ograniczonym oknie czasowym**
- **Dostarczenie wiarygodnych danych (jakość danych)**
- **Efektywność przetwarzania ETL**
- **Ewolucja źródeł danych**

Charakterystyka źródeł danych

- Różni producenci/technologie
- Różna funkcjonalność
  - bazy danych / nie bazy danych
  - dialekty SQL
  - sposoby dostępu i przetwarzania danych
- Różne modele danych
  - hierarchiczne, sieciowe
  - relacyjne
  - obiektowe
  - obiektowo-relacyjne
  - wielowymiarowe
  - XML

Charakterystyka źródeł danych

- Konflikty na poziomie struktur danych
  - różne reprezentacje danych (struktury)
- Konflikty na poziomie danych
  - Zduplikowane dane
  - Brakujące i błędne dane
  - Błędy wprowadzania wartości

- Hernandez M.A.; Stolfo S.J.: Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem. Data Mining and Knowledge Discovery 2(1):9-37, 1998
- Lee M.L.; Lu H.; Ling T.W.; Ko Y.T.: *Cleansing Data for Mining and Warehousing*. DEXA, 1999
- Kimball R., Caserta J.: The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning. Wiley, 2004
- Simitsis A., Vassiliadis P., Sellis T.K.: Extraction-Transformation-Loading Processes. Encyclopedia of Database Technologies and Applications 2005
- Rahm E., Do H.H.: Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin, (23):4, 2000

Różne reprezentacje danych

- Różne modele danych w źródłach (relacyjny, obiektowy, semistrukturalny)
- Różne typy danych
  - smallint, int, biging, decimal (SQLServer)
  - smallint, int, biging, float, real, double (DB2)
  - number, binary_integer (Oracle)
  - znakowe typy danych o stałej i zmiennej długości
- Różne ograniczenia integralnościowe
- Inna reprezentacja tych samych danych
  - Pracownicy{NIP, imię, nazwisko, adres_koresp}
  - Prac{NIP, imię_nazw, ulica, dom, kod, miasto}

Różne reprezentacje danych

- Homonimy
  - Produkty.kod – oznacza kod produktu
  - Klienci.kod – oznacza kod pocztowy
- Synonimy
  - Pacjenci.pesel
  - Pacjenci.pacjentID (z wartością peselu)

Konflikty na poziomie danych

- Różne ziarno agregacji
  - sprzedaż dzienna
  - sprzedaż tygodniowa
- Różne jednostki miary
  - cena {PLN, EUR, USD}
  - waga {kg, dkg}

Konflikty na poziomie danych

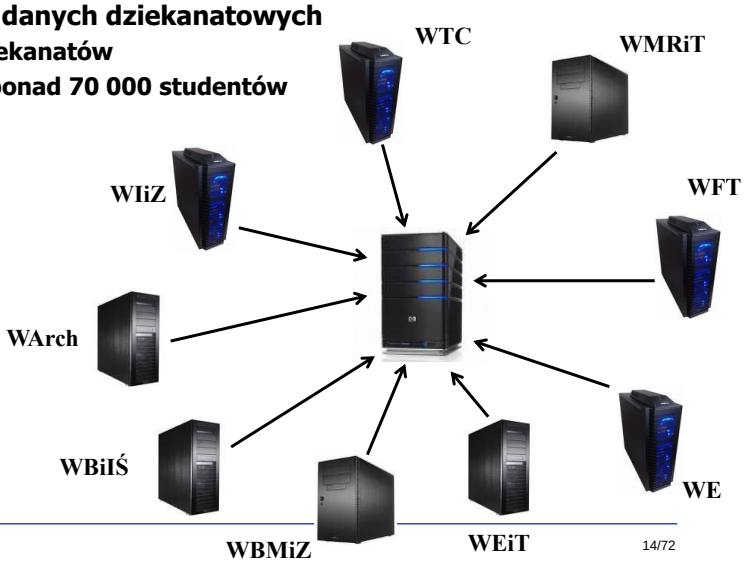
Atrybut	niedozwolona wartość	data_ur=30.13.1970	
	brakująca wartość	NIP=null	
	błąd literowy	Pozań	
	oznaczenia symboliczne	LC3X	czerwony metalik
	skróty	Pozn., P-ń, PŃ	
	wielkość liter	Poznań, poznań, POZNAŃ	
	różne oznaczenia symboliczne	płeć: {0, 1}, {K, M}, {kobieta, mężczyzna}	
	format	20-03-2008, 03/20/08	
	wartości złożone	R. Wrembel, 25.06.68, Szamotuły	
	kolejność wartości	{R.Wrembel} {Wrembel R.}	
	jednoznaczność	Wenecja (Włochy), Wenecja (Polska)	

Konflikty na poziomie danych

Rekord	niespełniona zależność pomiędzy wartościami atrybutów	cena_netto=100 cena_brutto=190	cena_brutto=cena_netto* 1,22
		ulica='Piotrowo' kod=62-300	
	naruszenie unikalności	{Robert Wrembel, WL8539024} {Bartosz Bębel, WL8539024}	nr dowodu osobistego powinien być unikalny
	wskazanie do nieistniejącego rekordu	{Robert Wrembel, Z20}	zespół Z20 nie istnieje
	duplikaty	{Robert Wrembel} {Wrembel Robert} {R. Wrembel}	
	sprzeczne wartości	{R. Wrembel, Szamotuły} {R. Wrembel, Poznań}	

Jakość danych - studium przypadku

- **Integracja danych dziekanatowych**
  - **9 BD dziekanatów**
  - **łącznie ponad 70 000 studentów**



Dane zduplikowane

- **Atrybut `st_czynny` (stan studenta, przyjmuje wartości ze słownika stanów studentów): atrybut zduplikowany, brak ograniczenia FK do tabeli słownikowej**
- **Atrybut `st_stan_aktualny` (przyjmuje wartości ze słownika stanów studentów): zdefiniowane ograniczenie FK do tabeli słownikowej**

czynny wg kolumny	czynny wg akt. stanu	liczba
tak	tak	17536
nie	tak	1
nie	nie	42988

Wartości unikalne

- **Nr indeksu**
  - **wartość unikalna w ramach jednej BD dziekanatu**
  - **wartość nieunikalna globalnie w ramach BD różnych dziekanatów ⇨ problem integracji danych**
  - **zidentyfikowano 49 par studentów o takim samym numerze indeksu, studenci w parze są innymi osobami**

Jakość danych nr indeksów

- **Postać nr indeksu: liczba + opcjonalnie litera**
  - **a - absolwent**
  - **d - drugi kierunek**
  - **s - skreślony lub zrezygnował**
  - **i - inne przypadki**

zawartość numeru albumu	liczba
tylko cyfry	65501
litera d na końcu	1247
litera a na końcu	433
litera s na końcu	992
litera i na końcu	982
litera p na końcu	603
litera x na końcu	358
litera g na końcu	146
litera r na początku	0
litera s na początku	225
wszystkie	71970
z literą na końcu	5333
z literą na początku	310
z literą	5625
białe znaki	293
inne znaki niż liczby	6471

Jakość danych PESEL

- **Różne osoby posiadające ten sam nr PESEL (również z tego samego wydziału)**
  - **13 powtarzających się numerów PESEL**
- **Poprawność wartości PESEL**

PESEL	liczba
wszystkie	71970
poprawny	56344
zła długość	8906
niepoprawne znaki przy dobrej długości	24
błąd sumy kontrolnej	6696
zła płeć przy założeniu M	763
zła płeć przy założeniu K	17

Jakość imion

- **Wykorzystano słownik imion (ponad 1700 imion;** http://piotr.eldora.pl/bazy-danych-kody-pocztowe-imiona-panstwa**)**

imię	liczba
wszystkie	86947
poprawne	86256
niepoprawne względem słownika	691
znaki inne niż litery	309
białe znaki	224
cyfry	57
białe znaki na początku lub końcu	16
dwa imiona	97
informacja o drugim kierunku	55
informacja o ITS	21
wielkimi literami	3
zła płeć przy imieniu M	30
zła płeć przy imieniu K	458

Jakość imion

imię	opis
Agata Agnieszka	wprowadzone dwa imiona
Natalia, Anna	wprowadzone dwa imiona z przecinkiem
Sławomir	zastąpienie polskiej litery diakrytycznej zwykłą
Joannna	literówka — nadmiarowa litera
Krzystof	literówka — brakująca litera
Przemysałw	literówka — przestawienie liter
Sikorska	nazwisko zamiast imienia
_Maciej	spacja występująca na początku lub końcu imienia
Marcin'	znak inny niż litera
wz.st.	informacja o wznowieniu studiów zamiast imienia
ITS	informacja o indywidualnym toku studiów zamiast imienia
-	brak imienia
Jakub 2 Kier.	informacja o drugim kierunku
Marcin S.Kazimierza	informacja o ojcu
Gniewosław	imię nie występuje w zbiorze poprawnych imion polskich, ale może być prawidłowe
Kevin	imię zagraniczne, może być prawidłowe

Jakość nazwisk

- **Wykorzystano słownik nazwisk (r.męski, 20000 najpopularniejszych nazwisk;** http://www.futrega.org/etc/nazwiska.html**)**

nazwisko	liczba
wszystkie	56607
niepoprawne względem słownika	11657
znaki inne niż litery, myślnik lub białe znaki	2
białe znaki	18
cyfry	1
białe znaki na początku lub końcu	2
wielkimi literami	1

Jakość danych słownikowych

- **Kategorie studiów**
  - **poprawne: {stacjonarne, niestacjonarne}**

kategoria studiów	liczba
Stacjonarne	9
Niestacjonarne	10
Trzeciego stopnia	7
Studia doktoranckie	1
Studia Wieczorowe	1
Niestacjonarne(wiecz)	1
Wieczorowe	1
Niestacjonarne(wieczorowe)	1

Jakość danych słownikowych

- **Rodzaje studiów**
  - **poprawne: {I stopnia, II stopnia, III stopnia}**
- **Kierunki studiów**

kierunek studiów	liczba
wszystkie	299
informacja o kategorii	137
informacja o rodzaju	154
informacja o miejscu	57
znaki inne niż litery i białe znaki	165
inaczej niż tylko pierwsza litera wielka	244

rodzaj studiów	liczba
Niestacjonarne I stopnia	11
Stacjonarne I stopnia	8
Niestacjonarne II stopnia	7
Stacjonarne II stopnia	7
Stacjonarne magisterskie	7
Niestacjonarne III stopnia	2
Niestacjonarne magisterskie	2
stacjonarne I stopnia	2
Stacjonarne III stopnia	2
Wieczorowe inżynierskie	2
Dzienne magisterskie uzupełniające	1
Niestacjonarne I stopnia (4 letnie)	1
Niestacjonarne II stopnia (uzupełniające)	1
Niestacjonarne II-stopnia	1
Niestacjonarne I-stopnia	1
Niestacjonarne magisterskie jednolite	1
Niestacjonarne uzupełniające	1
Stacjonarne I stopnia.	1
Stacjonarne II stopnia.	1
stacjonarne II stopnia	1
Stacjonarne II stopnia - 1,5-roczne	1
stacjonarne magisterskie	1
Stacjonarne magisterskie - 2-letnie	1
Stacjonarne magisterskie jednolite	1
Stacjonarne uzupełniające	1
Studia niestacjonarne	1
Studia stacjonarne	1
Wieczorowe	1
wieczorowe inżynierskie	1
Wieczorowe magisterskie uzupełniające	1
Zaoczne uzupełniające studia magisterskie	1
Zaoczne zawodowe	1

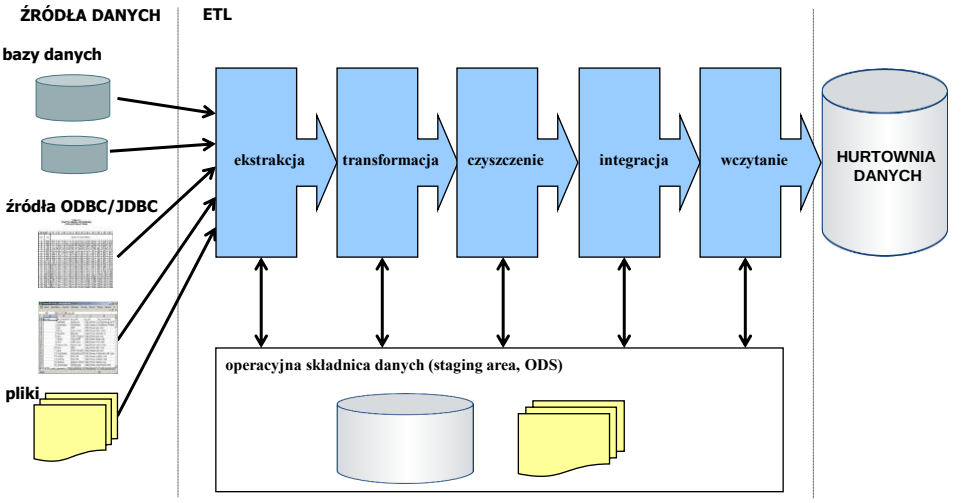
Jakość danych słownikowych

- **Słownik miast**

miejscowość	liczba
wszystkie	14293
znaki inne niż litery, myślnik lub białe znaki	567
cyfry	39
białe znaki na początku lub końcu	4
inaczej niż tylko pierwsza litera wyrazu wielka	11563

miejscowość	liczba baz	liczba adresów
POZNAŃ	9	12633
Poznań	2	369
?????POZNAŃ???	1	1
POZNAN	1	73
KONIN	9	1455
Konin	2	95
konin	1	23
KONIN 2	1	0
KALISZ	9	1748
Kalisz	2	76
KaLISZ	1	39
kalisz	1	21
KALisz	1	6
KALISz	1	2
KalISZ	1	2
kALISZ	1	2
kAlIsz	1	2
kaLISZ	1	2
kaliSZ	1	2
KALIsz	1	1
KAlisz	1	1
KaLiSZ	1	1
KaLiSz	1	1
KaLisz	1	1
kAlisZ	1	1
GNIEZNO	9	1517
Gniezno	3	45
PIŁA	9	1231
Piła	2	30
LESZNO	9	798
Leszno	2	25
WRZEŚNIA	9	732
Września	2	35
SWARZĘDZ	9	690
Swarzędz	2	33
BYDGOSZCZ	9	605
Bydgoszcz	2	17
BYDGOSKIE	1	1
BYDGOSZCZ ADRES	1	0
GORZÓW WLKP.	7	173
Gorzów Wielkopolski	7	92
GORZÓW WIELKOPOLSKI	4	358
GORZÓW WLKP	1	33
GORZÓW WLKP-	1	5

Architektura ETL

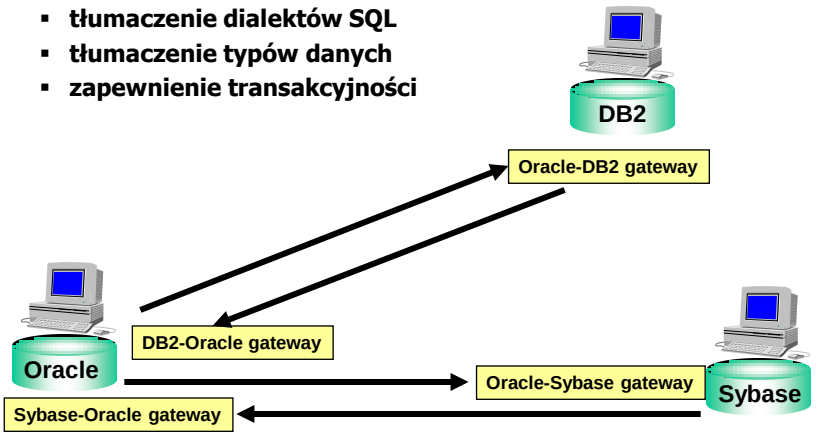


Źródła danych

- Zidentyfikowanie źródeł danych do zasilania HD
- Opis każdego źródła
  - obszar działalności (kadry, płace, marketing, ...)
  - rodzaj aplikacji obsługujących
  - ważność
  - użytkownik biznesowy (departamenty)
  - właściciel biznesowy
  - właściciel techniczny
  - SZBD
  - sprzęt + SO
  - liczba użytkowników/dzień
  - rozmiar bd
  - złożoność (liczba tabel, schemat)
  - liczba transakcji/dzień

Technologie dostępu

- **Gateway**
  - **tłumaczenie dialektów SQL**
  - **tłumaczenie typów danych**
  - **zapewnienie transakcyjności**



Technologie dostępu

ODBC/JDBC

- **standard definiujący metody dostępu do baz danych, bez względu na technologię implementacyjną tych baz danych**
- **ujednolicone metody dostępu implementowane w warstwie pośredniej pomiędzy aplikacją, a bazą danych ➔ sterownik ODBC/JDBC**
- **sterownik ODBC/JDBC ➔ interfejs programistyczny API**

- **OLE DB (Object Linking and Embedding DataBase)**
  - **API opracowane przez Microsoft, umożliwiające dostęp do różnych źródeł danych**
  - **dostęp do baz danych (standard ODBC)**
  - **dostęp do innych źródeł danych**
- **Sterowniki dostępu do plików tekstowych i XML**

Oprogramowanie

- Sybase EnterpriseConnect Data Access
  - dostęp do MS SQL Server, IBM DB2, Oracle i Informix (teraz IBM)
- IBM DataJoiner
  - dostęp do db Oracle, Sybase, Microsoft, Informix
- Oracle Transparent Gateways
  - dostęp do bd IBM DB2, Sybase Adaptative Server Enterprise , MS SQL Server
- Hyperion Integration Server
  - dostęp do bd IBM DB2, Sybase Adaptative Server Enterprise , MS SQL Server, Oracle

Wykrywanie zmian w źródłach

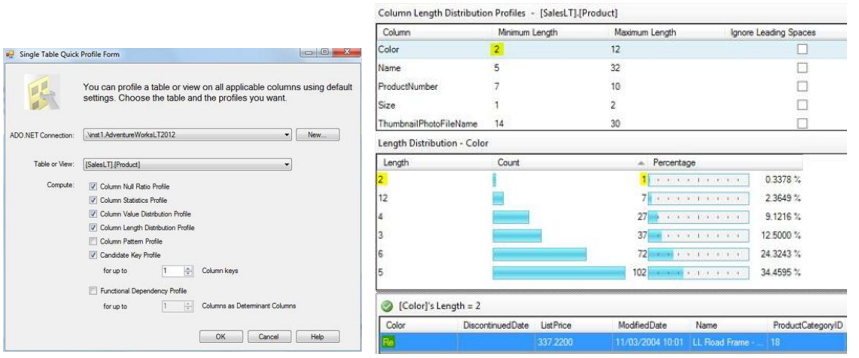
- Wymagania
  - minimalna ingerencja w oprogramowanie źródła
  - minimalny wpływ na pracę źródła
- Rozwiązania
  - kolumny audytu
    - w tabeli, data i czas operacji, rodzaj operacji
    - wypełnianie: wyzwalacze lub aplikacje
  - dziennik operacji w bazie danych (snapshot log)
  - analiza zawartości redo log
    - okresowo (log scraping)/ na bieżąco (log sniffing)
  - porównanie poprzedniego obrazu źródła z bieżącym
    - niska efektywność

Analiza źródeł danych

- Metody analityczne (statystyczne, eksploracja danych), których zadaniem jest określenie charakterystyki danych (data profiling)
- Metody analityczne
  - jakość danych
    - zidentyfikowanie kolumn z wartościami NULL/NOT NULL
    - liczba rekordów z wartościami pustymi lub domyślnymi dla każdego atrybutu (wart. domyślna może oznaczać brak wart. rzeczywistej)
    - zidentyfikowanie kolumn z wartościami unikalnymi
    - dozwolone długości atrybutów/wartości
    - dozwolone zakresy/zbiory wartości dla atrybutów
    - MIN, MAX, średnia, wariancja, odchylenie stand.
    - liczba rekordów z wartościami innymi niż dozwolone
    - krotność i rozkład wartości atrybutu
    - formaty wartości (np. daty, numery telef.)
    - zidentyfikowanie niepoprawnych wartości

Analiza źródeł danych

- Metody analityczne
  - struktura i zawartość źródła danych
  - dzienny przyrost liczby rekordów
- MigrationArchitect(Evoke Software), Integrity (Vality)
- SQL Server



Analiza źródeł danych

- Informatica Data Explorer
- Oracle
  - Data quality
  - Data profiling

Analiza źródeł danych

- Metody eksploracji danych ⇨ reguły asocjacyjne + wiedza dziedzinowa
  - Sapia C., Höfling G., et. al.: On Supporting the Data Warehouse Design by Data Mining Techniques
  - odkrywanie znaczenia atrybutów
    - (kraj='GB' → ki=2) wsparcie 95%: ki=kierownica; 2=strona prawa
  - uzupełnianie wartości pustych na podstawie reguł o wysokim wsparciu
  - zastępowanie wartości błędnych poprawnymi
  - identyfikowanie zależności funkcyjnych ⇨ odkrywanie kluczy potencjalnych
  - odkrywanie reguł biznesowych zdefiniowanych w aplikacjach
- WizRule (WizSoft), DataMiningSuite (InformationDiscovery)

Transformacja

- **Wymagania**
  - **Proces interakcyjny i iteracyjny**
    - **określenie kryteriów dopasowania + uruchomienie procesu + weryfikacja wyników + zmodyfikowanie kryteriów dopasowania**
  - **Proces rozszerzalny i łatwy do modyfikowania**
  - **Optymalizowalny**
  - **Automatyzacja max. liczby kroków**
  - **Minimalizacja danych do manualnej weryfikacji**

Transformacja

- **Transformacja do wspólnego modelu danych**
  - **{obiektowy, O-R, semistrukturalny, ...} ⇨ relacyjny**
- **Transformacja do wspólnej reprezentacji**
  - **Pracownik{NIP, imię, nazwisko, ulica, dom, kod, miasto}**
- **Usuwanie niepotrzebnych kolumn**
- **Często wymagana interwencja użytkownika**

Czyszczenie

- Ekstrakcja pól z ciągów znaków
  - ul. Piotrowo 2, 60-965 Poznań
  - układanie pól w kolejności
- Usuwanie wartości pustych
- Zamiana wartości błędnych na poprawne
  - słowniki ortograficzne
  - słowniki nazw (kraje, miasta, kody adresowe)
- Standaryzacja wartości
  - formatowanie wartości (np. daty)
  - przeliczanie walut
  - małe-duże litery
  - jednolite skróty
  - słowniki synonimów (Word Net)
  - słowniki skrótów

Czyszczenie

- Scalanie semantycznie identycznych rekordów
- Generowanie sztucznych identyfikatorów

Pesel	Imię	Nazwisko	Adres	Wykształcenie
55032206644	Robert	Wrembel	ul. Karkonoska	wyższe

NIP	Imię	Nazwisko	Ulica	Miasto	Kod	Wykształcenie
111-111-11-11	Robert	Wrembel	ul. Karkonoska 33	Pobiedziska	44-044	

ID	NIP	Pesel	Imię	Nazwisko	Ulica	Miasto	Kod	Wykształcenie
1	111-111-11-11	55032206644	Robert	Wrembel	ul. Karkonoska 33	Pobiedziska	44-044	wyższe

- IdCentric (FirstLogic), Trillium (TrilliumSoftware)

Integrowanie - eliminowanie duplikatów

- Porównywane rekordy muszą być oczyszczone
  - usunięte znaki specjalne, interpunkcyjne
  - rozwinięte skróty
- Rekordy różnią się nieznacznie wartościami
  - {Wrembel, Robert, ul. Wyspiańskiego, Poznań}
    {Wrębel, Robert, ul. Wyspiańskiego, Poznań}
- Porównanie identyfikatorów naturalnych (np. nr dowodu, paszportu, silnika}
- Brak identyfikatorów naturalnych
  - sortowanie + porównanie sąsiednich n rekordów (okno o szerokości n)
  - funkcja podobieństwa (np. nazwiska i adresy identyczne)
    - wagi podobieństwa dla różnych atrybutów
  - przybliżone łączenie (approximate join)

Eliminowanie duplikatów

- Prosta miara podobieństwa
  - liczba pasujących atomowych łańcuchów / całkowita liczba atomowych łańcuchów w dopasowywanych ciągach
  - Polit. Pozn., Wydz. Inf. i Zarządzania; Instytut Informatyki,
  - Politechnika Poznańska, Inst. Infrom.
  - miara=4/11

Eliminowanie duplikatów

- Soundex
  - algorytm grupowania nazw zgodnie z ich wymową
  - nazwy wymawiane tak samo (mimo innej pisowni) posiadają tę samą wartość Soundex
  - soundex('Smith')=soundex('Smit')=S530
- Dystans Levenhsteina (Levenhstein/edit distance)
  - miara podobieństwa dwóch łańcuchów znaków źródłowego L1 i docelowego L2
  - dystans mierzony minimalną liczbą operacji wstawiania i usuwania (i modyfikowania) znaków w łańcuchu prowadzących do uzyskania L2 z L1
  - L1 i L2 identyczne ⇨ dystans=0
  - ABC ⇨ ABCDEF: dystans=3
  - DEFCAB ⇨ ABC: dystans=5
- Merge (Sagent), DataCleanser (EDD)

Zasilanie HD

- Kiedy odświeżać
  - synchronicznie (po zatwierdzeniu transakcji w źródle) ⇨ HD czasu rzeczywistego
  - asynchronicznie ⇨ tradycyjne HD
    - z zadaną częstotliwością
    - na żądanie
- Co przesyłać
  - dane (Oracle)
  - transakcje (Sybase, SQL Server)
- Jak odświeżać
  - w sposób przyrostowy
  - w sposób pełny
- Jak często odświeżać
  - odświeżanie wsadowe
  - odświeżanie strumieniowe (near real-time DW)

Zasilanie HD - efektywność

- W ściśle określonym oknie czasowym
  - wczytanie danych o rozmiarze około 5TB - przynajmniej 8h
- Odczyt tylko danych potrzebnych
- Unikać
  - DISTINCT, operatorów zbiorowych,
  - NOT i połączeń nierównościowych (zwykle wymagają full scan)
  - funkcji w klauzuli WHERE
  - GROUP BY w zapytaniu pobierającym dane ze źródła
    - sortowanie w systemie źródłowym (niska efektywność)
    - interakcja z przetwarzaniem OLTP w źródle
  - wyzwalaczy w HD
- Pobieranie danych ze źródeł
  - źródło ma indeksy na atrybutach klauzuli WHERE i dobry optymalizator ⇨ zapytanie pobierające dane z WHERE
  - brak indeksów lub słaby optymalizator ⇨ pobrać wszystko i odfiltrować dane w ETL (filtry ETL)

Zasilanie HD - efektywność

- Oddzielenie operacji UPDATE od INSERT
  - Uwaga: UPDATE nie jest wspierany ścieżką bezpośrednią
  - zastąpienie UPDATE przez DELETE i INSERT
  - liczba UPDATE > INSERT => TRUNCATE TABLE + INSERT
- Indeksy
  - usunięcie + utworzenie ⇔ modyfikowanie na bieżąco
  - indeksy w przypadku UPDATE
    - usunięcie indeksów niewykorzystywanych do optymalizacji UPDATE
    - wykonanie operacji UPDATE
    - usunięcie pozostałych indeksów
    - wstawienie rekordów
    - utworzenie indeksów
- Ograniczenia integralnościowe
  - wyłączyć przed wczytywaniem

Zasilanie HD - efektywność

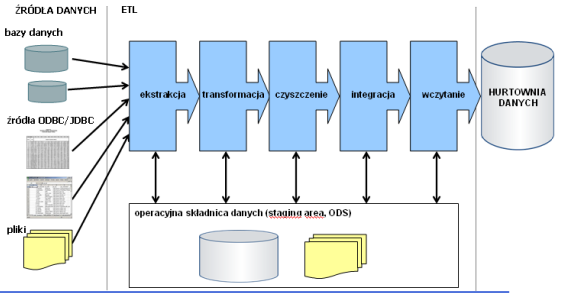
- Redo log
  - wyłączenie zapisów do redo log
    - dane wstawiane przez oprogramowanie ETL zarządzające również wycofywaniem nieudanych operacji
    - dane wstawiane wsadowo
    - możliwość powtórzenia nieudanych wstawień
  - wyłączenie zapisów do redo log dla tabeli
- Ścieżka bezpośrednia (direct load path)
- Filtrowanie danych z plików w systemie operacyjnym (polecenie awk)
- Sortowanie i obliczanie agregatów w silniku ETL (nie w bazie danych)
- Sortowanie w systemie operacyjnym (polecenie sort)

Zasilanie HD - efektywność

- Transformacja
  - jeśli można to unikać operacji w bazie danych - stosować przepływ zadań (workflow)
- Wczytywanie równoległe do wielu partycji
- Stosować natywne sterowniki do źródeł danych (unikać ODBC)
- Zebranie statystyk po zasileniu
- Defragmentacja bazy danych

Cel stosowania ODS

- Odseparowanie przetwarzania ETL od operacyjnych źródeł danych
  - niedostępna dla użytkowników źródeł i HD
- Zapewnienie możliwości powtórzenia przerwanego/wycofanego procesu ETL bez potrzeby sięgania do źródeł danych
- Dane źródłowe i częściowo przetworzone

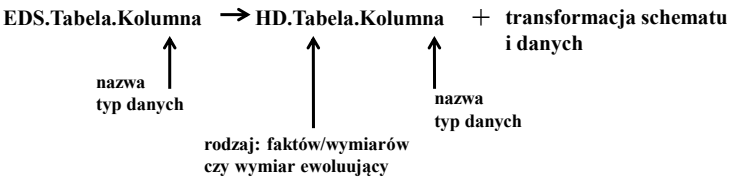


Zawartość ODS

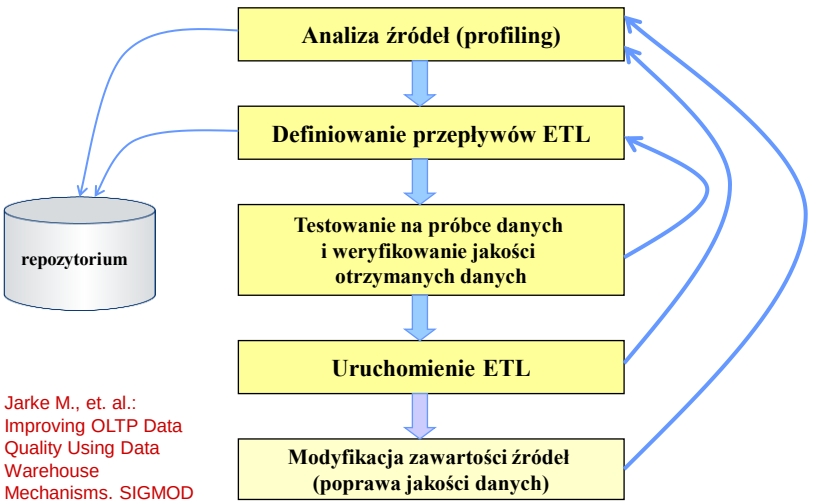
- Dane ze źródeł
  - elementarne
  - zagregowane
  - tabele i pliki
- Tabele odwzorowań kluczy (ODS ⇨ EDS)
  - zastosowanie w lineage

Odwzorowanie danych

- Rejestrowanie pochodzenia obiektów (rekordów) w HD (lineage)
  - obiekty źródłowe
  - operacje aplikowane do obiektów źródłowych przez ETL
  - rekordy w HD posiadają atrybuty przechowujące identyfikatory rekordów źródłowych, z których powstały
- Odwzorowanie transformacji danych źródłowych w dane w HD



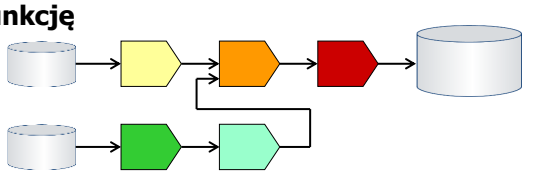
Projektowanie ETL



- Jarke M., et. al.: Improving OLTP Data Quality Using Data Warehouse Mechanisms. SIGMOD Record, (28):2, 1999

Implementacja ETL

- ETL - przepływ pracy (workflow) zbudowany z sekwencji transformacji
- Transformacja - realizuje funkcję
  - agregacja
  - filtrowanie
  - łączenie
  - normalizowanie
  - pobranie rekordu z innej tabeli (lookup)
  - generowanie numerów
  - sortowanie
  - adaptery źródeł danych
  - modyfikowanie danych
  - interfejs XML
  - definiowana przez projektanta



Metadane ETL

- Biznesowe
  - słowniki pojęć biznesowych
  - odwzorowania pojęć biznesowych w obiekty HD
  - reguły biznesowe
  - jakość danych
- Sterujące wykonaniem ETL
  - harmonogramy
  - skrypty
  - logi z wykonania
  - monitorowanie
- Techniczne

Metadane ETL

- Metadane techniczne
  - **opis źródeł** (lokalizacja, struktura, zawartość)
    - rodzaj źródła (relacyjna bd, obiektowa bd, xml, html, arkusz kalkulacyjny, ...)
    - struktura/schemat
    - metody dostępu
    - użytkownicy i prawa dostępu
    - wyniki analizy (profilowania) źródeł
    - dzienny przyrost danych
    - rozmiary danych
    - przyrost danych w czasie (np. dzienny)
    - średnia długość wiersza
  - **opis HD**
    - schemat
    - struktury fizyczne
    - statystyki dot. danych

Metadane ETL

- Metadane techniczne
  - organizacja przestrzeni dyskowej ODS i HD
  - charakterystyki danych zasilających (gotowy zbiór zasilający)
  - statystyki dla optymalizacji
  - implementacje algorytmów (transformacje, czyszczenie, eliminowanie duplikatów)
  - słowniki transformacji (np. nazwy miast)
  - techniki odświeżania (pełne/przyrostowe, okresy)
  - statystyki dot. odświeżania (liczba rekordów przesłanych, rekordy błędne)
  - nazwy zadań ETL korzystające z danej struktury

Metadane ETL

- Opisujące procesy ETL
  - struktura przepływu pracy
  - odwzorowania źródło ⇔ HD
  - odwzorowania rekordów źródłowych w docelowe (lineage)
  - definicje transformacji (nazwa, realizowany cel, wejście, wyjście, algorytm)
  - skrypty i zadania (nazwa, realizowany cel, źródło, struktury docelowe, pliki logów, pliki sterujące, statystyki efektywnościowe z wykonania, obsługa wyjątków/awarii)
  - harmonogram uruchamiania ETL (częstotliwość, obsługa wyjątków/awarii, pliki logów, statystyki efektywnościowe z wykonania)
  - logi z pracy ETL
  - charakterystyka danych
  - fizyczna organizacja przestrzeni dyskowej

Wymagania dla ETL

- Efektywność
  - zakończenie w z góry zadanym czasie
- Niezawodność
  - restart po zatrzymaniu na skutek błędów
  - odtwarzanie po awarii
- Zarządzanie
  - określanie częstotliwości odświeżania
  - automatyczne startowanie
    - czasowe
    - token - informacja przesłana ze źródła (plik, wpis w tabeli) o dostępności danych ze źródła
  - wycofywanie i restartowanie zadań od początku
  - wstrzymywanie i startowanie zadań
- Zapewnienie jakości danych
  - poprawność wartości i struktury

Wymagania dla ETL

- Bezpieczeństwo
  - na skutek awarii
  - autoryzacja dostępu
- Predefiniowane operatory/operacje
  - reguły transformacji struktury, danych i czyszczenia specyfikowane deklaratywnie
- Automatyczne generowanie kodu
- Łatwość modyfikowania
- Możliwość dołączania własnych programów
- Uruchamianie wsadowo
- Automatyczne raportowanie o zakończeniu, błędach, wyjątkach i postępie
- Możliwość wykonania równoległego
- Wykorzystanie metadanych

Wymagania dla ETL

- Pobieranie danych ze źródeł
  - często najbardziej czasochłonne
- Szacowanie czasu wykonania
- Monitorowanie
  - czas procesora
  - RAM
  - przepustowość
  - konflikty w dostępie do dysków

Oprogramowanie ETL

- Gotowe
  - szybsza realizacja procesów ETL
  - zintegrowane repozytoria danych
  - zarządzanie metadanymi
  - szeregowanie procesów
  - wbudowane sterowniki do wielu systemów
  - analiza zależności pomiędzy komponentami
  - inkrementalne odświeżanie
  - równoległość operacji
- Programowane
  - koszt wytworzenia i testowania oprogramowania
  - dedykowane do jednego rozwiązania

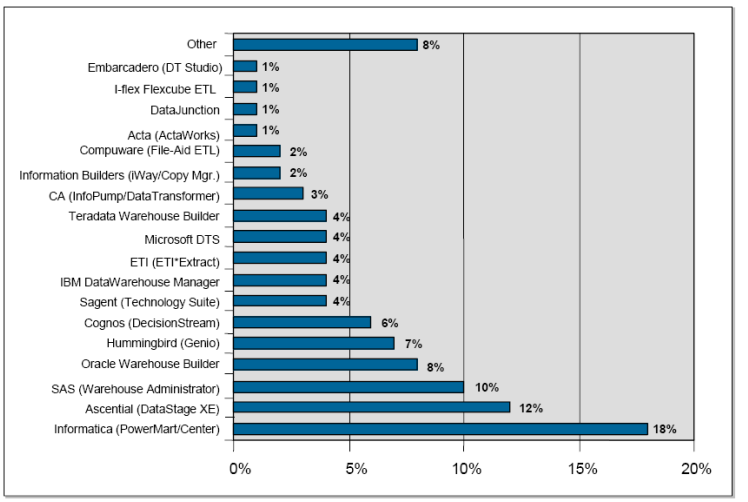
Komponenty oprogramowania ETL

- Ekstrakcja
  - interfejsy dostępu do źródeł
  - profilowanie źródeł
  - wykrywanie zmian (change data capture)
  - pobieranie danych ze źródeł
- Czyszczenie
  - uspójnianie i uzupełnianie danych (miary jakości)
  - obsługa i logowanie błędów
  - eliminowanie duplikatów
- Zasilanie
  - zarządzanie wymiarami
  - zarządzanie faktami
  - generatory sztucznych ID
  - wyliczanie agregatów

Komponenty oprogramowania ETL

- Zarządzanie przepływem pracy ETL
  - automatyczne uruchamianie zadań
  - odtwarzanie po awarii ETL
  - lineage i zarządzanie zależnościami między obiektami
  - monitorowanie pracy
  - zrównoleglanie pracy
  - składnica metadanych

Systemy komercyjne



Source: Giga Information Group

Systemy prototypowe

- **AJAX - Inria**
  - Galhardas H., Florescu D., Shasha D., Simon E.: An Extensible Framework for Data Cleaning. ICDE, 2000
  - Galhardas H., Florescu D., Shasha D., Simon E.: AJAX: An Extensible Data Cleaning Tool. SIGMOD, 2000
- **Potter's Wheel - Berkeley**
  - Raman V., Hellerstein J.M.: Potter's Wheel: An Interactive Data Cleaning System. VLDB, 2001
- **Arktos II - National Univ. of Athens, Univ. of Ioannina**
  - Vassiliadis P., A. Simitsis, Georgantas P, Terrovitis M.: A Framework for the Design of ETL Scenarios. CAiSE, 2003
  - Simitsis A., Vassiliadis P., Skiadopoulos s., Sellis T.: Data Warehouse Refreshment. In Data Warehouses and OLAP: Concepts Architectures and Solutions. IGI, 2007
  - Simitsis A., Vassiliadis P., Sellis T.: Optimizing ETL processes in data warehouses. ICDE, 2005
  - Simitsis A., Vassiliadis P., Sellis T.: State-Space Optimization of ETL Workflows. IEEE TKDE (17):10, 2006
  - Tziovara V., Vassiliadis P., Simitsis A.: Deciding the physical implementation of ETL workflows. DOLAP, 2007

AJAX

- **Wejście: zbiór tabel z niespójnymi, błędnymi, zduplikowanym rekordami**
- **Wyjście: zbiór tabel z danymi spójnymi, poprawnymi, bez duplikatów**
- **Założenia**
  - **tabele posiadają klucze podstawowe**
  - **łączone tabele w związku 1:1**

AJAX - Komponenty

- Usługa transformacji danych
  - standaryzacja wartości
  - transformacja do innej postaci
  - MAPPING makro-operator

```
CREATE MAPPING MG1
SELECT k.klID, k.imie, nazwisko, ulica, miasto, kod,
       k.nrTel, wyksztalc
INTO Klienci_Era_S
FROM Klienci_Era k
LET nazwisko=INITCAP(k.nazwisko)
   [ulica, miasto, kod]=ExtractAdr(k.adres)
    wyksztalc=IF(k.wyksztalc is not null) THEN RETURN k.wyksztalc
              ELSE RETURN 'nieznane'
```


AJAX - Dopasowanie

- Usługa dopasowania rekordów
  - wyznaczenie zbioru pasujących do siebie rekordów
    - przygotowanie jednego zbioru rekordów z n źródeł
    - miara podobieństwa rekordów <0, 1>
  - przygotowanie zbioru rekordów bez duplikatów
  - MATCH makro-operator

```
CREATE MATCH MH1
FROM Klienci_Era ke, Klienci_Orange ko
LET sim1=nazwSimF(ke.nazwisko, ko.nazwisko)
    sim2=adresSimF(ke.adres, ko.adres)
SIMILARITY=IF (sim1>0.9 and sim2>0.8) THEN RETURN MIN(sim1,sim2)
           ELSE IF (sim1 between 0.6 and 0.89 and
                    sim2 between 0.7. and 0.8) THEN RETURN sim1
           ELSE RETURN 0
THRESHOLD SIMILARITY>=0.7
```

  - Wynik działania ⇨ tabela dopasowania
    - D {ID_Kl_Era, ID_Kl_Orange, współcz_dopasowania}

AJAX - Eliminowanie Duplikatów

- Usługa eliminowania duplikatów i konstruowania spójnego, zintegrowanego wyniku
  - manualne
  - półautomatyczne
  - automatyczne THRESHOLD > x

```
CREATE MAPPING MG2
SELECT id, nazwisko, adres, ... INTO Klienci
FROM MH1
LET id=IDGen(D.ID_Kl_Era, D.ID_Kl_Orange)
    sim1=nazwSimF(D.ID_Kl_Era.nazwisko, D.ID_Kl_Orange.ko.nazwisko)
    sim2=ulicaSimF(D.ID_Kl_Era.adres, D.ID_Kl_Orange.adres)
SIMILARITY
    nazwisko=IF (sim1>0.9) THEN RETURN D.ID_Kl_Era.nazwisko
    ulica=IF (sim2>0.9) THEN RETURN D.ID_Kl_Orange.ulica
    .....
    adres=CONCAT(ulica, miasto, kod)
THRESHOLD SIMILARITY>=0.89
```


Potter's Wheel

- Interaktywny i iteracyjny proces transformacji i czyszczenia danych
  - zestaw predefiniowanych transformacji
    - formatowanie wartości (Format)
    - scalenie n wartości (Merge)
    - rozbicie na n wartości (Split)
    - dodanie, skopiowanie, usunięcie kolumny (Add, Copy, Drop)
    - zamiana kolumn na wiersze (Fold)
    - zamiana wierszy na kolumny (Unforld)
  - transformacje aplikowane na bieżąco (interaktywnie) ⇨ wyniki widoczne natychmiast w aplikacji typu arkusz kalkulacyjny
  - transformacje aplikowane tylko do danych aktualnie widocznych w oknie
  - wskazanie rodzaju transformacji na przykładzie konkretnych wartości ⇨ system wnioskuje właściwą transformację
  - kompilacja sekwencji transformacji do C, Perl lub makro Potter's Wheel

Arktos II

- Model konceptualny przepływu ETL ⇨ odwzorowany w model implementacyjny
- Graficzna reprezentacja modelu konceptualnego
  - koncept - reprezentuje źródłową lub docelową strukturę danych
    - złożony z atrybutów
  - transformacja
    - filtrowanie danych
    - transportowanie danych (ftp, szyfrowanie, kompresowanie)
    - projekcja
    - agregacja
    - połączenie
    - operacje na zbiorach
    - funkcja użytkownika
    - sortowanie

Arktos II

- Graficzna reprezentacja modelu konceptualnego
  - ograniczenie integralnościowe ETL - nakładane na obiekty przepływu ETL
  - związki
    - jest częścią - wiązanie atrybutów z konceptami
    - kandydat (wykorzystywany we wczesnych fazach projektowania) - wskazywanie potencjalnych źródeł danych dla przepływu ETL i potencjalnych struktur docelowych w HD
    - aktywny kandydat - wskazanie wybranego źródła/struktury docelowej
  - odwzorowanie źródło-struktury docelowe

- Simitsis A., Vassiliadis P., Sellis T.: Optimizing ETL processes in data warehouses. ICDE, 2005
- Simitsis A., Vassiliadis P., Sellis T.: State-Space Optimization of ETL Workflows. IEEE TKDE (17):10, 2006
- Tziovara V., Vassiliadis P., Simitsis A.: Deciding the physical implementation of ETL workflows. DOLAP, 2007