POZNAN UNIVERSITY OF TECHNOLOGY

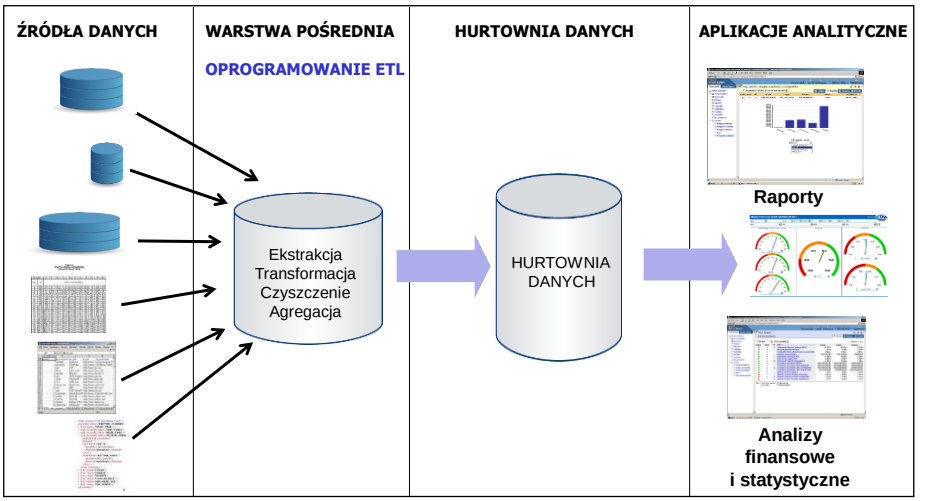
# Technologie Zasilania i Odświeżania Hurtowni Danych

**Robert Wrembel**
**Poznan University of Technology**
**Institute of Computing Science**
**Poznań, Poland**
**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**

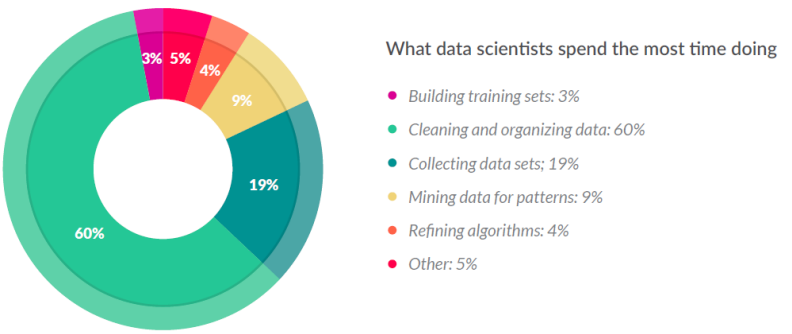
# Charakterystyka ETL

- **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**



# Charakterystyka ETL

➲ **Data Science Report. 2016, Crowd Flower**

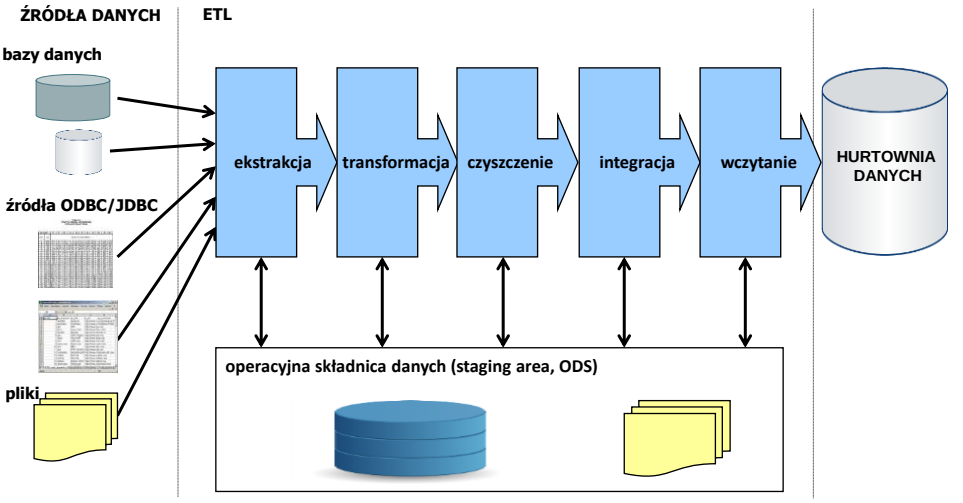


# Wyzwania

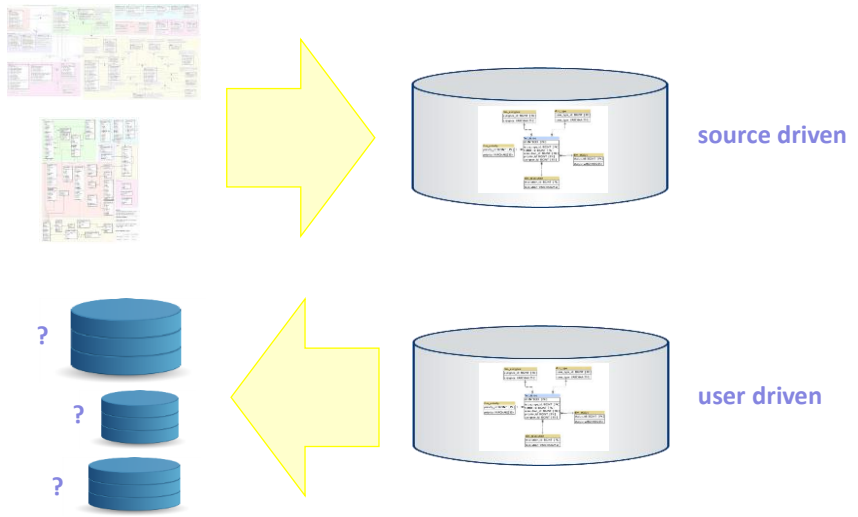
- Przetwarzanie dużych wolumenów danych w ograniczonym oknie czasowym
- Dostarczenie wiarygodnych danych (jakość danych)
- Efektywność przetwarzania ETL
- Ewolucja źródeł danych



# Architektura ETL



# Projekt HD






# Projekt HD

- Analiza zawartości źródeł danych
- Technologie dostępu
- Profilowanie
- Odczyt danych
  - pełny
  - zmian
- Transformacja
- Czyszczenie / uspójnianie
- Eliminowanie duplikatów
- Wczytanie do HD

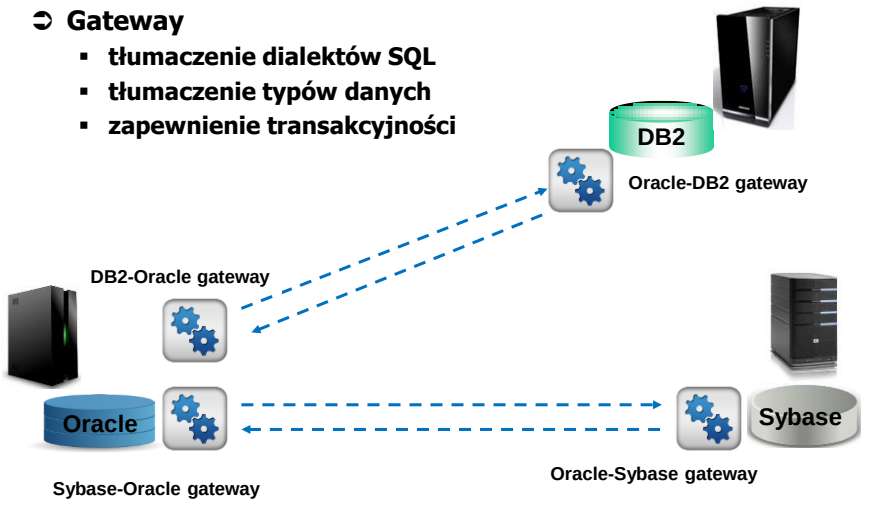
# Źródła danych

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



# 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 Enterprise Connect Data Access**

- **dostęp do MS SQL Server, IBM DB2, Oracle, źródła ODBC**

➲ **IBM InfoSphere Federation Server**

- **dostęp do Oracle, Sybase, Microsoft**

➲ **Oracle Database Gateways**

- **dostęp do IBM DB2, Sybase Adaptative Server Enterprise , MS SQL Server, Teradata, Informix**

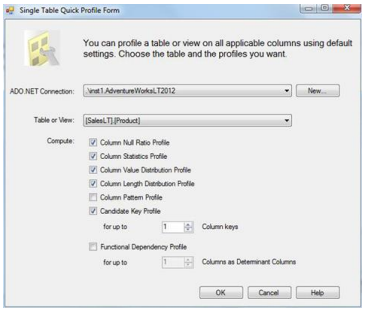
# Analiza źródeł danych

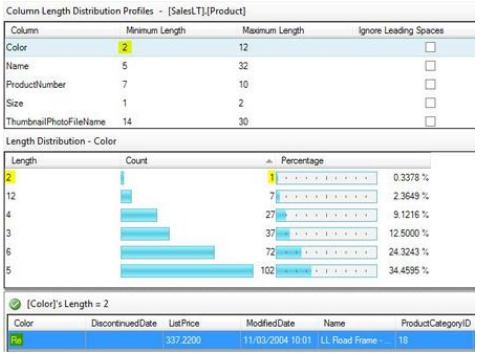
- **Data profiling ⇨ metody analityczne (statystyczne, eksploracja danych), których zadaniem jest określenie charakterystyk źródeł danych**
- **Jakość danych**
  - **zidentyfikowanie kolumn z wartościami NULL/NOT NULL**
  - **liczba rekordów z wartościami NULL lub domyślnymi dla każdego atrybutu**
  - **zidentyfikowanie kolumn UNIQUE**
  - **dozwolone długości wartości atrybutów**
  - **dziedziny atrybutów**
  - **MIN, MAX, średnia, wariancja, odchylenie standardowe**
  - **krotność i rozkład wartości atrybutu**
  - **liczba rekordów z wartościami innymi niż spodziewane**
  - **formaty wartości (np. daty, numery telef.)**
  - **zidentyfikowanie niepoprawnych wartości**



# Analiza źródeł danych

- **Struktura i zawartość źródła danych**
- **Dzienny przyrost liczby rekordów**
- **Oprogramowanie**
  - **MigrationArchitect(Evoke Software)**
  - **Integrity (Vality)**
  - **SQL Server**



# Analiza źródeł danych

- **Oprogramowanie**
  - **Informatica Data Explorer**
  - **Oracle Data quality i Data profiling**

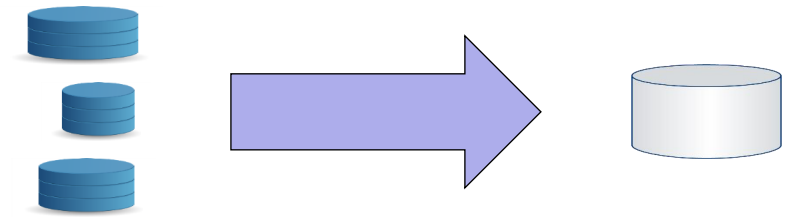



# Analiza źródeł danych

- **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)**

# Loading data into DW

- Reading the whole data source
  - text/binary dump files ⇨ DB export
  - XML files
  - SQL select + gateway / ODBC
  - snapshots
- Reading changes
  - need to detect data changes





# Detecting data changes

- Requirements
  - minimum or none source system changes
  - minimum interference with a data source
- Solutions
  - audit columns
  - snapshot comparison
  - system maintained log of changes on a table (e.g., snapshot log)
  - snapshots
  - triggers ⇨ synchronous transfer
  - analysis of redo log (transaction log)
    - periodically (log scraping)
    - on-line - continuously (log sniffing)

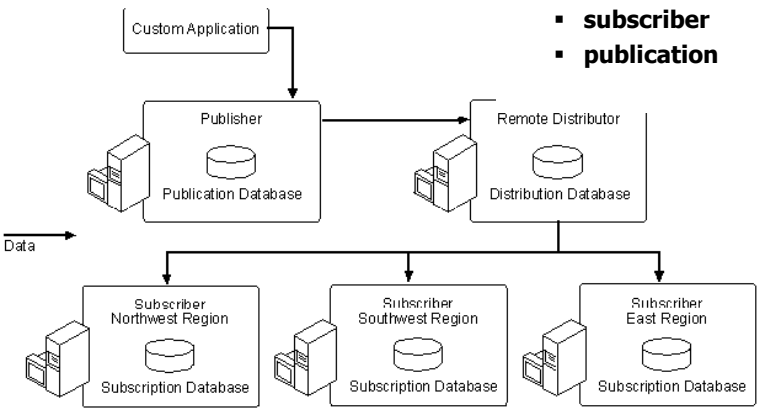
# Snapshot/replica

- Copy of a table or the subset of its columns and rows
- Refreshing
  - automatic with a defined interval
  - on demand
- SQL Server
- IBM DB2
- Oracle



# SQL Server

- Actors (nodes)
  - publisher
  - distributor
  - subscriber
  - publication

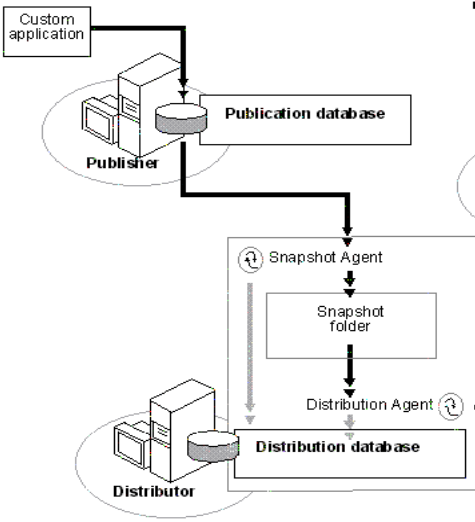


# SQL Server

- **Article**
  - **database object to be replicated**
  - **can be**
    - **a table**
    - **a subset of table's columns**
    - **a subset of table's rows**
- **Subscriber**
  - **a node receiving publications**
- **Subscription**
  - **contains publications**
  - **defines replication schedule**
  - **a set of subscribers**
  - **push / pull subscription**



# SQL Server - snapshot replication

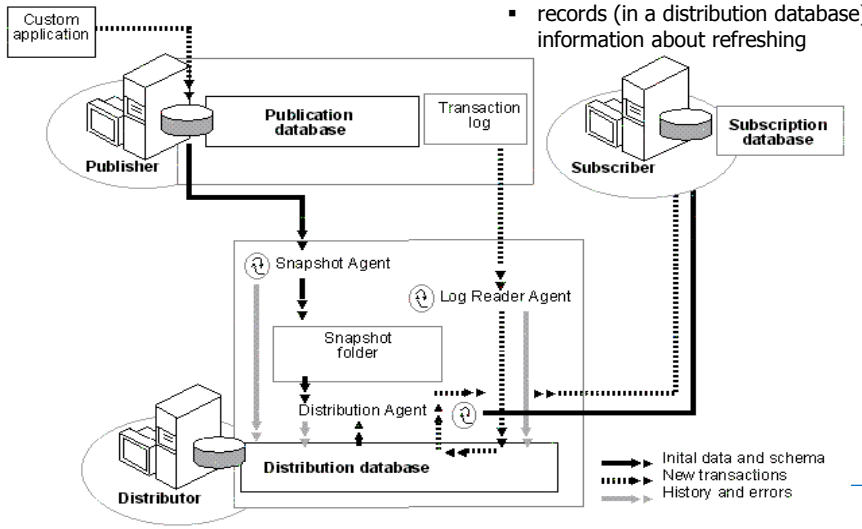



- Snapshot Agent
  - creates snapshot files
  - records (in a distribution database) information about refreshing
- Distribution Agent
  - sends a snapshot to a subscriber

Data or transactions
History and errors

# SQL Server - transactional replication

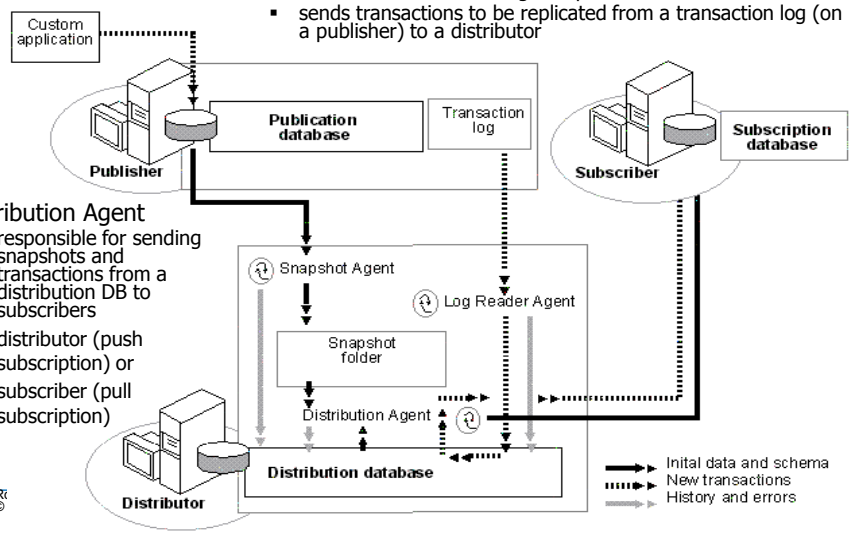
- Snapshot Agent
  - creates snapshot files
  - records (in a distribution database) information about refreshing




25
25

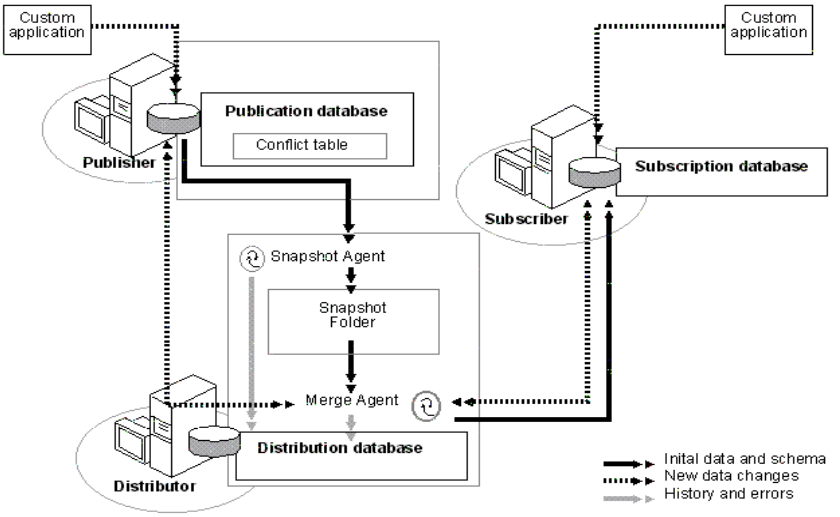
# SQL Server - transactional replication

- Log Reader Agent
  - monitors a transaction log of a publisher
  - sends transactions to be replicated from a transaction log (on a publisher) to a distributor




- Distribution Agent
  - responsible for sending snapshots and transactions from a distribution DB to subscribers
  - distributor (push subscription) or
  - subscriber (pull subscription)

# SQL Server - merge replication



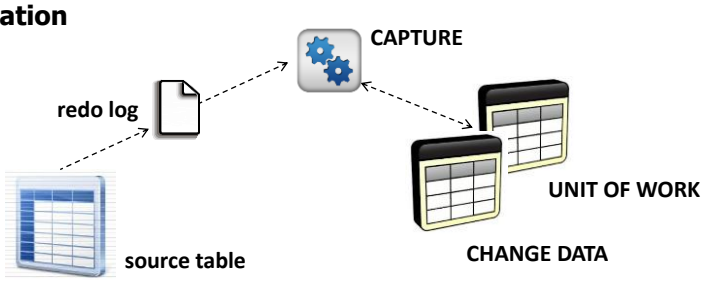



# SQL Server - replication summary

- **Snapshot replication**
  - **full refreshing**
  - **data shipping**
- **Transactional and merge replication**
  - **incremental refreshing**
  - **transaction shipping**
- **Automatic or on demand propagation**

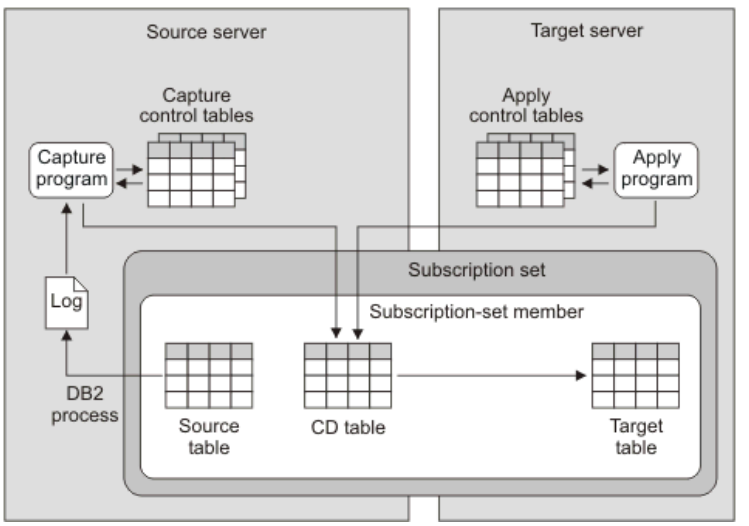
# IBM DB2

- Process CAPTURE ⇨ detecting changes in base tables
  - Each base table has associated table CHANGE DATA created by a system
  - Table UNIT OF WORK stores information on committed transactions
- Process APPLY ⇨ applying detected changes to replicas
- Process MONITOR ⇨ manages and monitors a replication

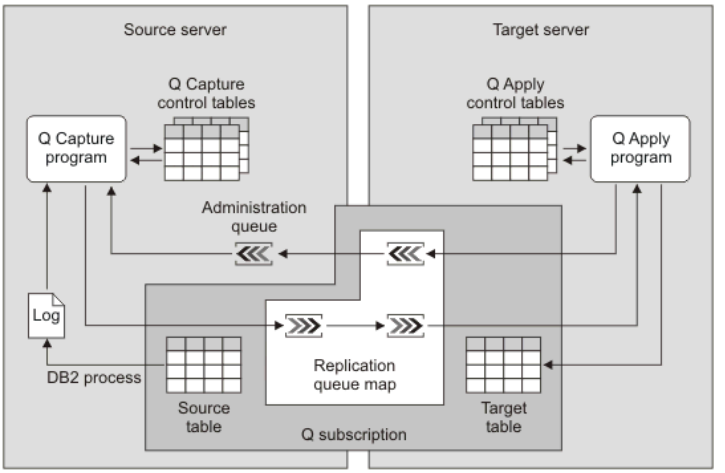





# IBM DB2



# IBM DB2 Queue replication




**IBM InfoSphere Replication Server**

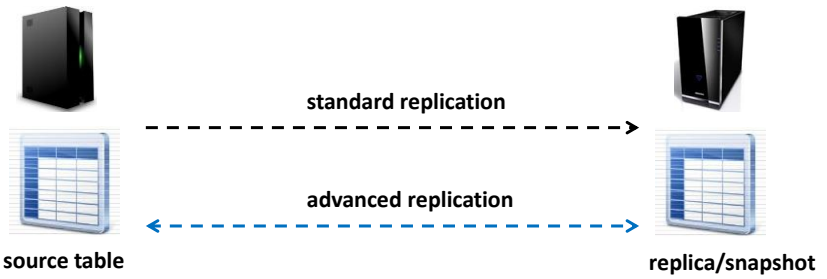


# IBM DB2 - replication summary

- **Incremental**
- **One-way (replica is read-only)**
- **Refreshing**
  - **automatic**
  - **automatic synchronous**
  - **event based**
- **Transaction shipping**
- **Queue replication**

# Oracle - snapshot replication

- Standard replication ⇨ replica is read-only
- Advanced replication ⇨ replica is updateable
  - single-master
  - multi-master






# Oracle - snapshot replication

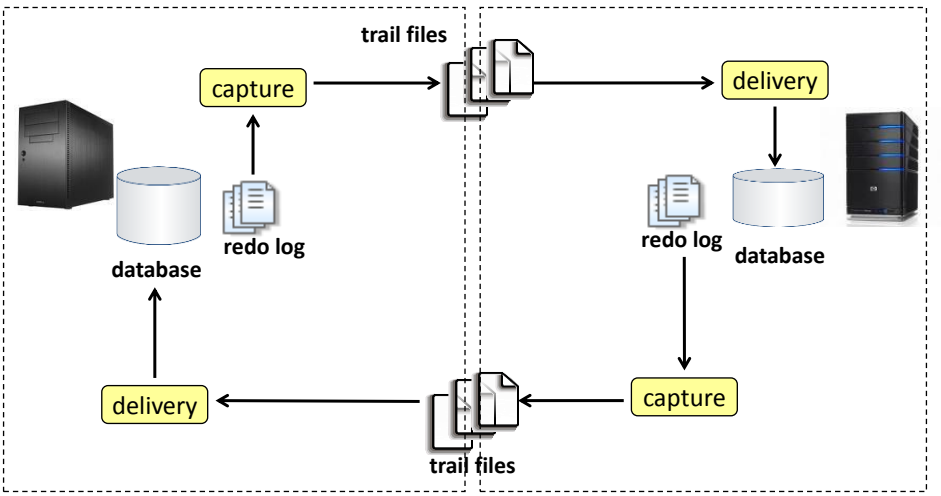
- Concept
  - a copy of a table (base table) stored in a remote DB
  - a refreshing process associated with a snapshot
- Implementation
  - table + index
- Definition (SQL)
  - refreshing method
    - full
    - incremental
    - incremental not always possible
  - refreshing moment
    - on demand
    - periodically within a defined time interval
  - row identification for incremental refreshing
    - primary key
    - ROWID
  - query

# Oracle - Golden Gate

- "Real-time" synchronization between nodes in DDBS
  - applicable to real-time/near real-time data synchronization
- Incremental replication → log sniffing
- Log-based capture from MS SQL Server, IBM DB2, MySQL
- Changes can be propagated to MS SQL Server, IBM DB2, MySQL
- Modules
  - capture
  - trail files
  - delivery



# Oracle - Golden Gate



# Oracle - Golden Gate

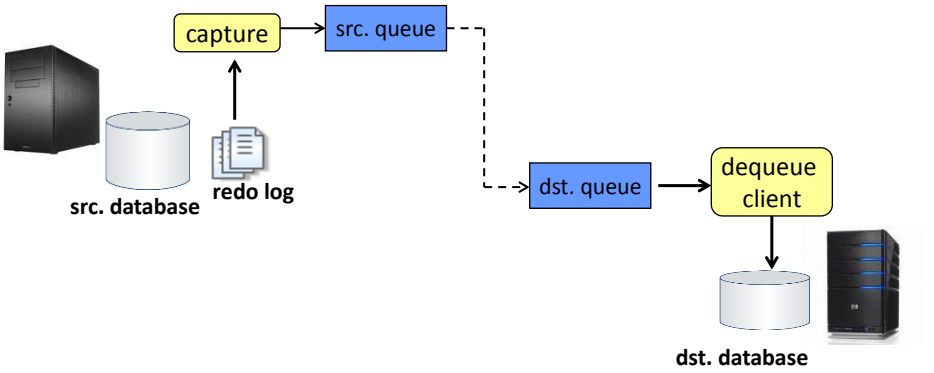
- Capture: resides on the source database
  - reads inserts, updates, and deletes
  - delivers only committed transactions
- Trail files: contain the database operations for the changed data in a transportable, platform-independent data format
  - reside on the source and/or target server outside of the database
  - Capture module reads once, and then immediately moves the captured data to the external trail file for delivery to the target
- Delivery: reads the content of a trail file and applies it to a target
  - native SQL is used for the target
  - target can also be any ODBC compatible data storage
  - the order of source transactions is preserved



# Oracle - Streams

- Process Capture: detects database events DDL and DML → redo log analysis
  - rules used by a capture process to determine which changes to capture
  - captured changes (messages) are stored in a queue
  - modes of capturing
    - from online redo log
    - from archived redo log
- Dequeue Client
  - can be either process Apply or a user application
  - rules determine which messages are dequeued
  - multiple destination databases can read from the same queue

# Oracle - Streams






# Oracle - Data Guard

- Standby database ⇨ for backup and recovery
- DB changes transmitted to a standby directly from memory
- Active Data Guard 11g
  - allows read-only access to a stand by while constantly synchronizing it with a primary DB

# Oracle Replication - Summary

- Snapshot replication
  - full or incremental
  - automatic or on-demand
  - data shiping
  - replica read only or updateable
- GoldenGate
  - incremental
  - transaction shipping
- Streams
  - incremental
  - data shipping
- Data Guard
  - incremental
  - transaction shipping



# Transformacja

- Wymagania
  - Proces interakcyjny i iteracyjny
    - określenie kryteriów transformacji + uruchomienie procesu + weryfikacja wyników + zmodyfikowanie kryteriów
  - 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

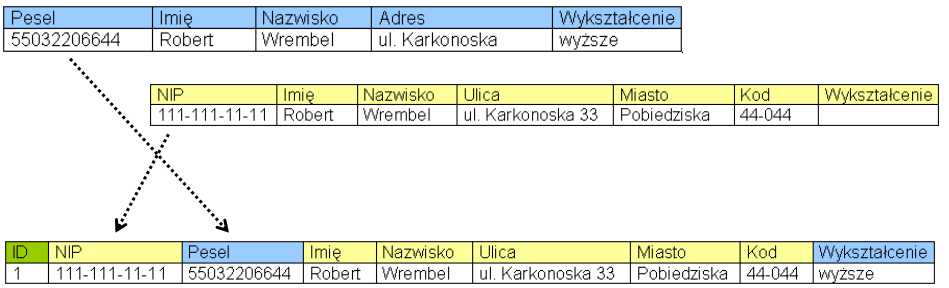


# Czyszczenie

- 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 tą 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 (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) ⇨ near real-time DW
  - asynchronicznie ⇨ tradycyjne HD
    - z zadaną częstotliwością
    - na żądanie
- Co przesyłać
  - dane (Oracle)
  - transakcje (Sybase, SQL Server)
- Jak odświeżać
  - w sposób przyrostowy
  - w sposób pełny
- Sposób zasilania
  - wsadowo ⇨ zastosowania typowe
  - strumieniowo ⇨ near real-time DW



# Zasilanie HD - efektywność

- W ściśle określonym oknie czasowym
  - wczytanie danych o rozmiarze około 5TB - może zająć około 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

# Zasilanie HD - efektywność

- **Pobieranie danych ze źródeł**
  - **selektywnie**
    - **źródło ma indeksy na atrybutach klauzuli WHERE i dobry optymalizator**
    - **zapytanie bardzo selektywne**
  - **pełen odczyt i filtrowanie w warstwie ETL**
    - **brak indeksów lub słaby optymalizator**
    - **zapytanie o niskiej selektywności**



# Zasilanie HD - efektywność

- **Oddzielenie operacji UPDATE od INSERT**
  - **Uwaga: UPDATE nie jest wspierany ścieżką bezpośrednią**
  - **zastąpienie UPDATE przez DELETE i INSERT**
  - **jeśli 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**

# Zasilanie HD - efektywność

- Ograniczenia integralnościowe
  - wyłączyć przed wczytywaniem
- Czy w HD muszą być ograniczenia integralnościowe?



# 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
- Sortowanie w systemie operacyjnym (polecenie sort)

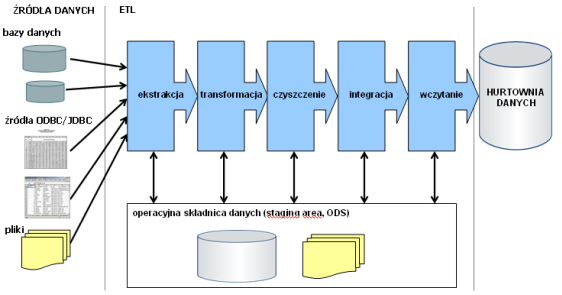
# Zasilanie HD - efektywność

- 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
- Data provenance



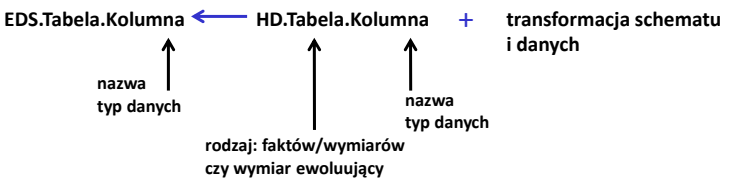
# Zawartość ODS

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

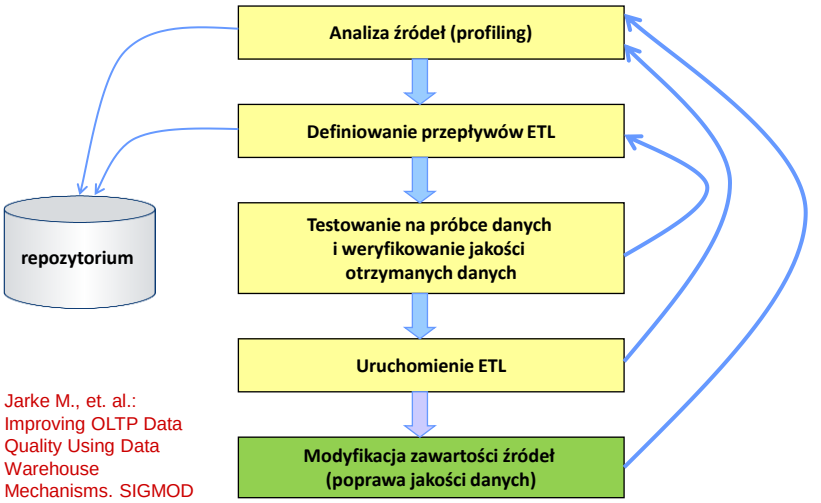


# Odwzorowanie danych

- Rejestrowanie pochodzenia obiektów (rekordów) w HD (data provenance)
  - 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

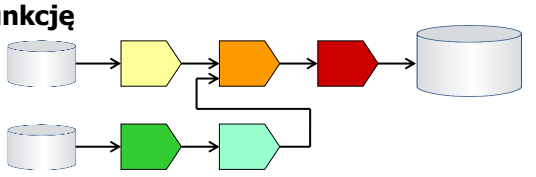



- Jarke M., et. al.: Improving OLTP Data Quality Using Data Warehouse Mechanisms. SIGMOD Record, (28):2, 1999

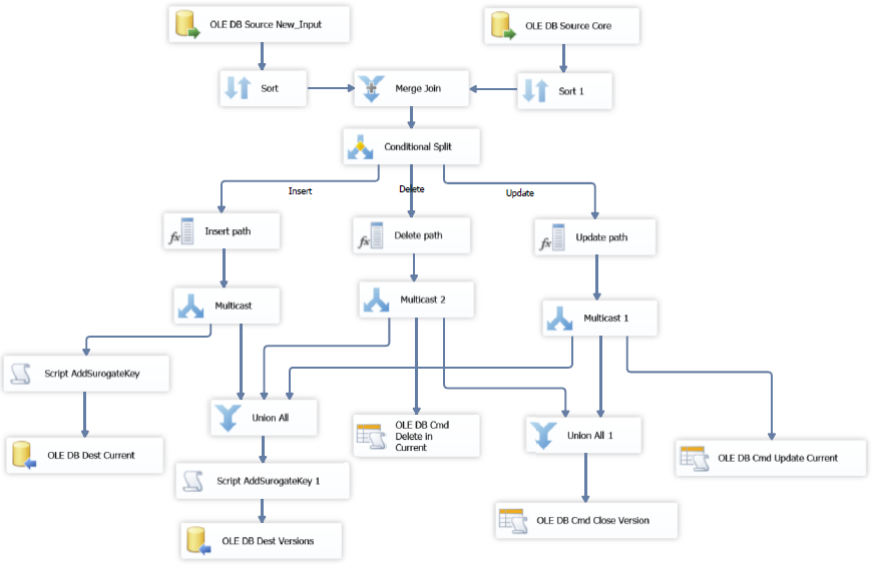


# Implementacja ETL

- **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**



# Implementacja ETL




# Metadane ETL

- **Biznesowe**
  - **słowniki pojęć biznesowych**
  - **odwzorowania pojęć biznesowych w obiekty HD**
  - **reguły biznesowe**
- **Techniczne**
  - **opisujące źródła danych**
  - **opisujące hurtownię danych**
  - **opisujące procesy ETL i sterujące nimi**

# 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**
    - **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**
  - **definicje transformacji (nazwa, realizowany cel, wejście, wyjście, algorytm)**
  - **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)
  - 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



# 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



# Wymagania dla ETL

- Możliwość wykonania równoległego
- Wykorzystanie metadanych
- Szacowanie czasu wykonania
- Monitorowanie
  - czas procesora
  - RAM
  - przepustowość
  - konflikty w dostępie do dysków
- Optymalizacja wykonania

# Oprogramowanie ETL

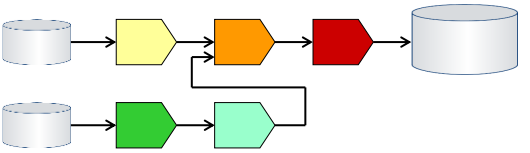
- Gotowe
  - szybsza realizacja procesów ETL
  - zintegrowane repozytoria danych
  - zarządzanie metadanymi
  - szeregowanie procesów
  - wbudowane sterowniki do wielu systemów
  - analiza zależności pomiędzy komponentami
  - inkrementalne odświeżanie
  - równoległość operacji
- Programowane
  - koszt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



# Off-the-shelf technology?

- Financial Times (18 Sep, 2013): Why big IT projects crash
- www.ft.com
- "... the Texas state auditor's office examined 13 IT projects, nine of which had overrun. It concluded, admittedly on a small sample, that agencies using commercial off-the-shelf technology exceeded their budgets by a smaller amount and took less time to complete their projects than those that did not"

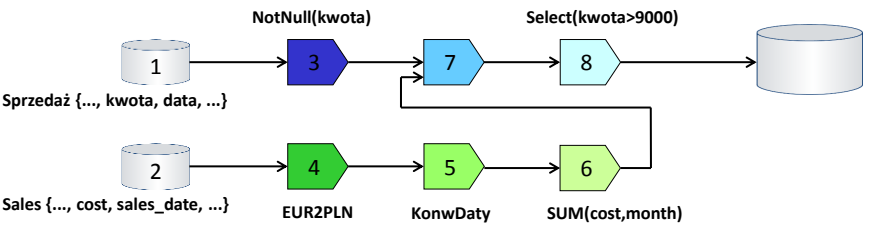
# Optymalizacja ETL



- 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

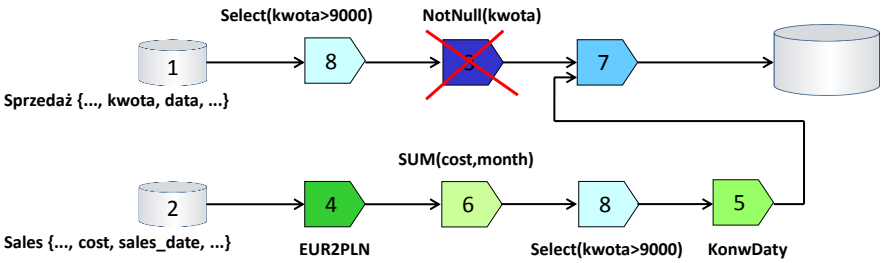


# Optymalizacja ETL




- 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

# Optymalizacja ETL




- 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



# Komponenty oprogramowania ETL

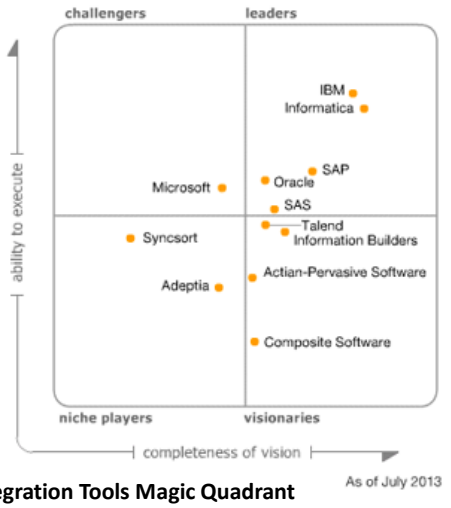
- Ekstrakcja
  - interfejsy dostępu do źródeł
  - profilowanie źródeł
  - wykrywanie zmian (change data capture)
  - pobieranie danych ze źródeł
- Czyszczenie
  - uspójnianie i uzupełnianie danych (miary jakości)
  - obsługa i logowanie błędów
  - eliminowanie duplikatów
- Zasilanie
  - zarządzanie wymiarami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



# Gartner Report




**Gartner's 2013 Data Integration Tools Magic Quadrant**