Wprowadzenie do Hurtowni Danych

Mariusz Rafało
mrafalo@sgh.waw.pl

WPROWADZENIE DO SYSTEMÓW ZASILANIA HURTOWNI DANYCH

Narzędzia ETL

Extract - Transform – Load

- Pobieranie danych z systemów źródłowych
- Transformacja danych do postaci docelowej
- Ładowanie danych do systemów docelowych

Typowo używane podczas zasilania Hurtowni Danych

- Obecnie ewoluują do kompletnych platform integracji danych, wspierających wszystkie techniki integracji danych w różnych trybach
- Narzędzia ETL stanowią rozwiązanie komplementarne w stosunku do platform integracji aplikacji

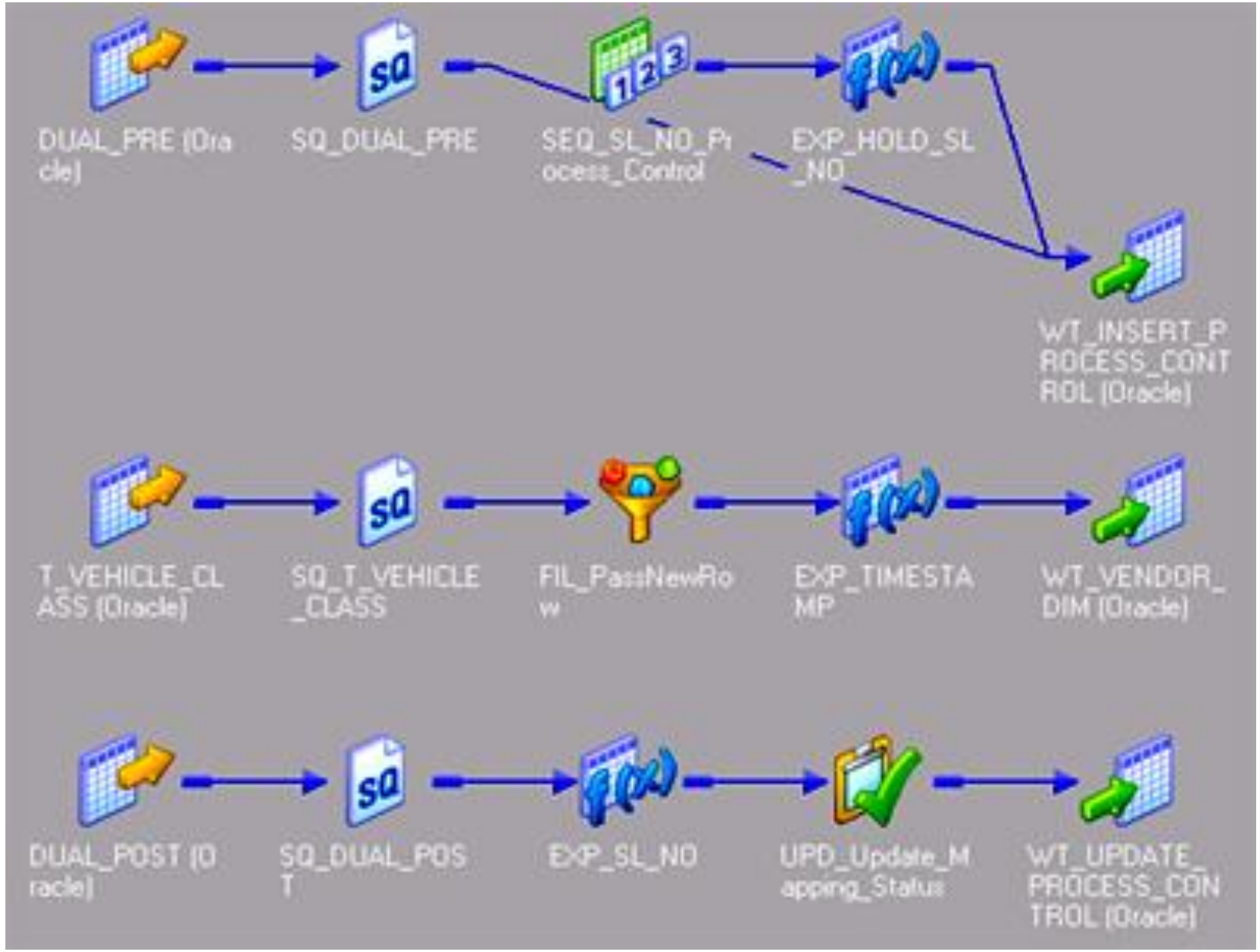
Historia ETL

Generatory kodu (początek lat 90)

- Generatory programów w językach trzeciej generacji (np. COBOL)
- Skompilowany kod działał na różnych platformach, nie wymagał odrębnego serwera
- Proste zadania ETL
- Słabo zautomatyzowane przetwarzanie, manualne zarządzanie kodem i harmonogramowaniem zadań.

Silniki transformacji (połowa lat 90)

- Interpretery wykonujące skrypty na serwerze ETL lub bazy danych
- Cały proces ETL odbywa się w silniku, a nie w systemach źródłowych
- Złożone zadania tworzone w środowisku graficznym
- Sekwencje zadań i ich harmonogramy zawarte w metadanych
- Równoległe przetwarzanie


DUAL_PRE (Oracle)
SQ_DUAL_PRE
SEQ_SL_NO_Process_Control
EXP_HOLD_SL_NO
WT_INSERT_PROCESS_CONTROL (Oracle)
T_VEHICLE_CLASS (Oracle)
SQ_T_VEHICLE_CLASS
FIL_PassNewRow
EXP_TIMESTAMP
WT_VENDOR_DIM (Oracle)
DUAL_POST (Oracle)
SQ_DUAL_POST
EXP_SL_NO
UPD_Update_Mapping_Status
WT_UPDATE_PROCESS_CONTROL (Oracle)

ETL – główne funkcjonalności

Obsługa różnych źródeł danych

- standardowe konektory do baz danych i plików płaskich
- możliwość tworzenia własnych adapterów
- obsługa Web Services i XML

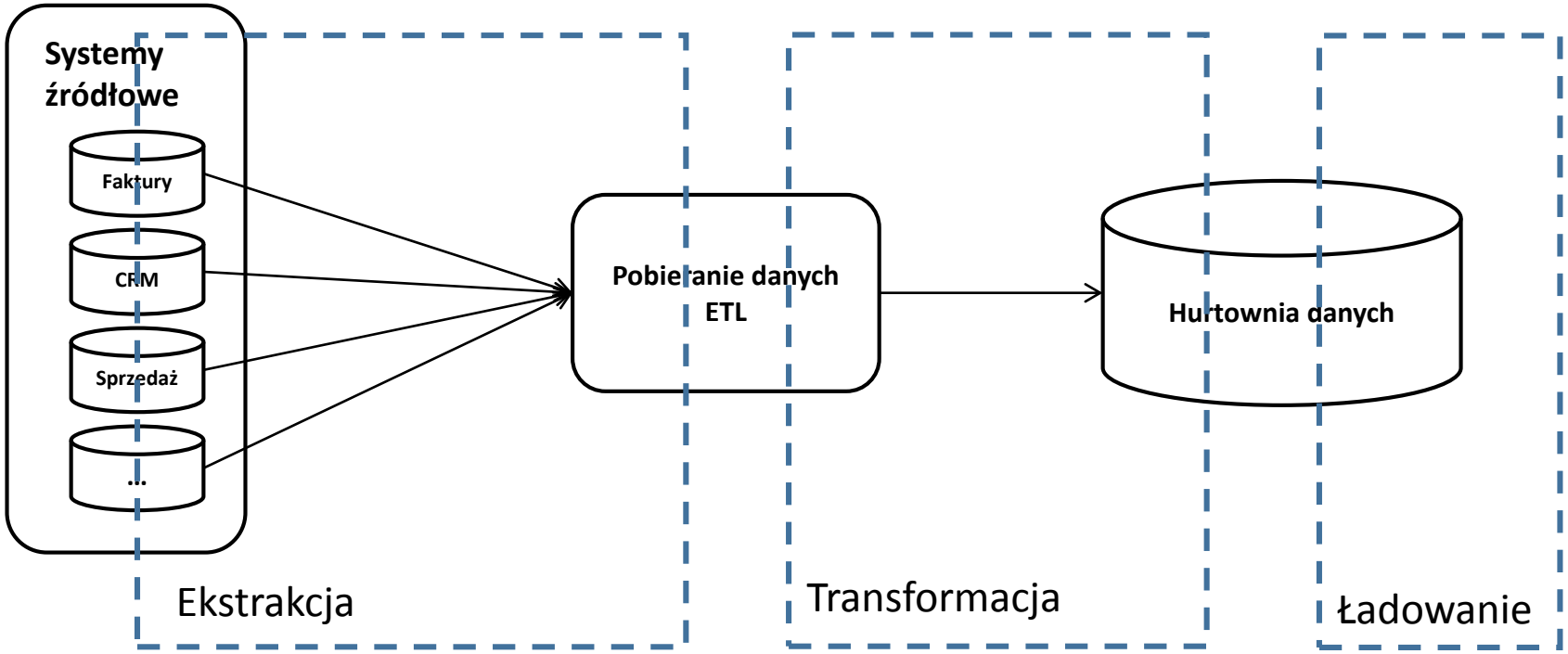
Możliwość realizowania złożonych transformacji

- wykorzystaniem gotowych komponentów
- możliwość budowania własnych komponentów
- graficzne projektowanie procesów przepływu danych

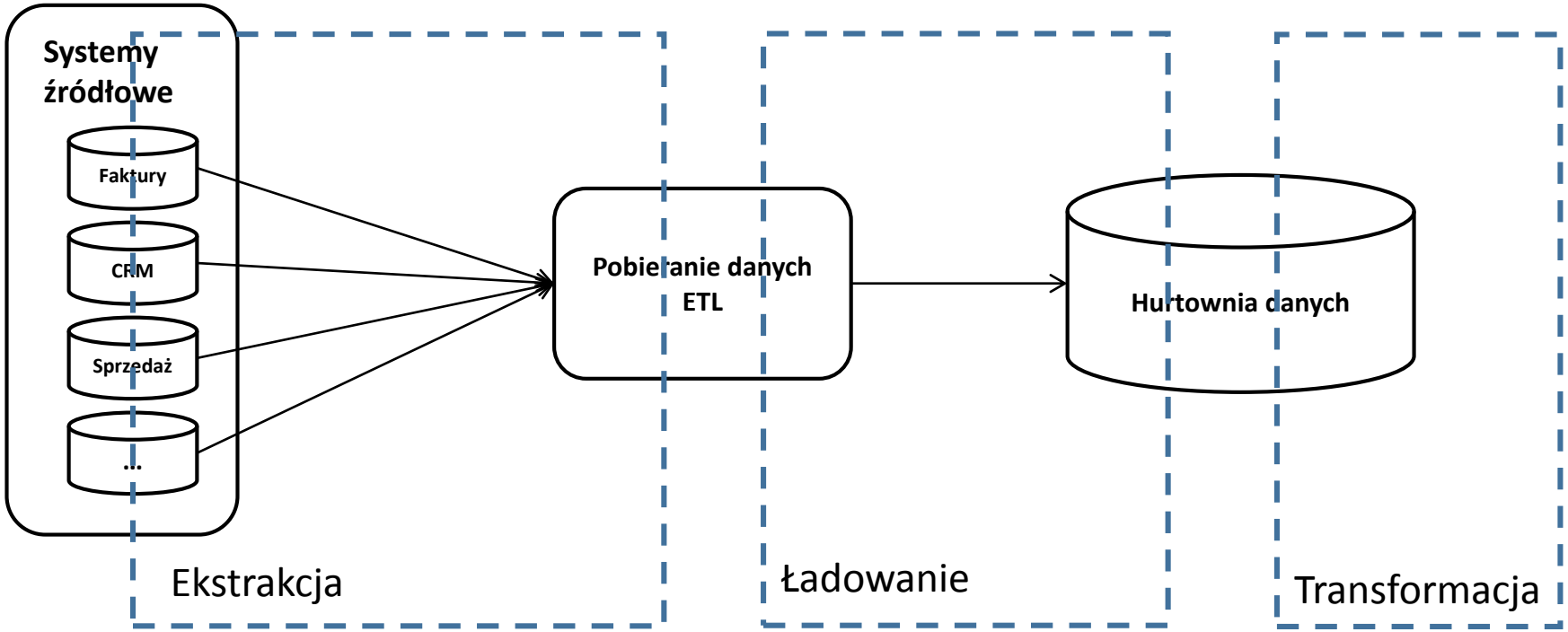
Możliwość elastycznego harmonogramowania przetwarzania

- harmonogram przetwarzania
- zależności pomiędzy procesami
- zdarzenia

ETL a ELT


Systemy źródłowe
Faktury
CRM
Sprzedaż
...
Pobieranie danych ETL
Hurtownia danych
Ekstrakcja
Transformacja
Ładowanie

ETL a ELT


Systemy źródłowe
Faktury
CRM
Sprzedaż
...
Pobieranie danych ETL
Hurtownia danych
Ekstrakcja
Ładowanie
Transformacja

ETL – dodatkowe możliwości ETL

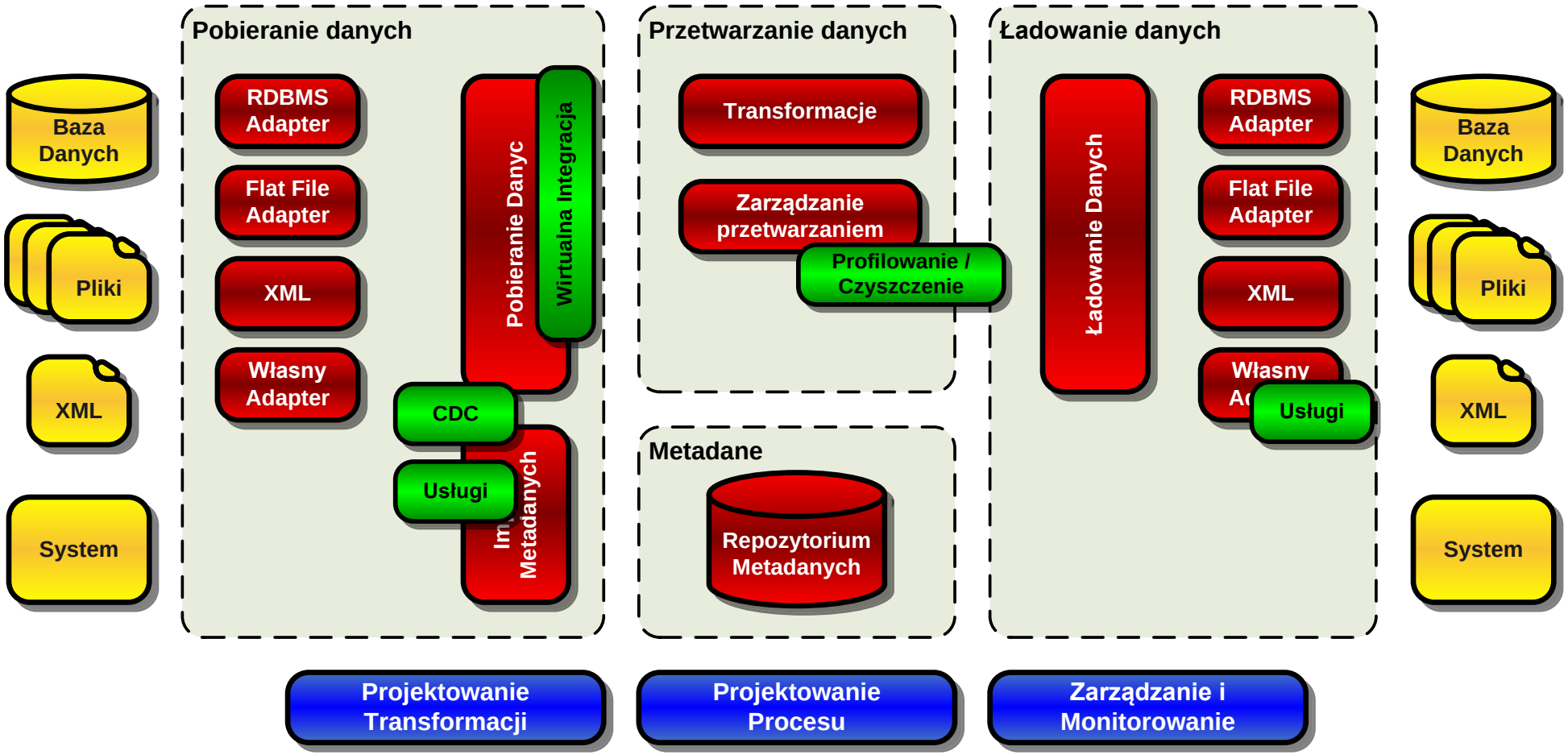
Wbudowane mechanizmy poprawy jakości danych:

- Czyszczenie
- Profilowanie
- Deduplikacja

Optymalizacja działania i skalowalność

- Równoległe przetwarzanie
- Architektura klastra
- Cache’owanie danych
- Load balancing – rozdzielenie przetwarzania tak, aby żadna maszyna nie była przeciążona

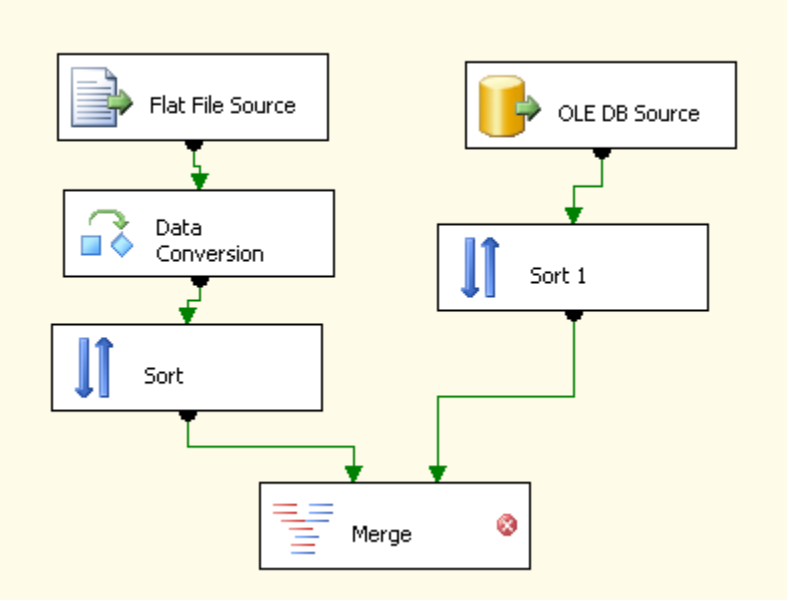
Architektura platformy ETL



Komercyjne platformy ETL

Zalety narzędzi ETL

- Dają możliwość zidentyfikowania rodowodu danych oraz przeprowadzenia analizy zależności
- Posiadają wbudowane konektory do różnego rodzaju baz danych
- Mają zoptymalizowane działanie w kontekście bardzo dużych zbiorów danych (paralelizm, wielowątkowość)
- Transformacje i proces przetwarzania nie są zaszyte w kodzie, tylko przedstawione na diagramach
- Mogą je wykorzystywać osoby biznesowo-techniczne, nie będące profesjonalnymi programistami
- Istnienie zintegrowanego repozytorium, które synchronizuje metadane systemów źródłowych, docelowych i narzędzi BI



PROJEKTOWANIE SYSTEMU ZASILANIA

Fazy projektowania systemu zasilania

1. opracowanie architektury systemu zasilania
2. selekcja źródeł danych
3. strategia pobierania danych
4. opracowanie transformacji
5. opracowanie harmonogramów wykonania
6. określenie sytuacji wyjątkowych i reakcji na nie

Opracowanie architektury systemu zasilania

Najważniejsze czynniki mające wpływ na planowanie architektury:

- ilość i zróżnicowanie źródeł danych
- wolumen przetwarzanych danych
- dostępność źródeł danych (okna czasowe)
- moc obliczeniowa systemów źródłowych
- przepustowość sieci

Fazy projektowania systemu zasilania

1. opracowanie architektury systemu zasilania
2. selekcja źródeł danych
3. strategia pobierania danych
4. opracowanie transformacji
5. opracowanie harmonogramów wykonania
6. określenie sytuacji wyjątkowych i reakcji na nie

Identyfikacja źródeł

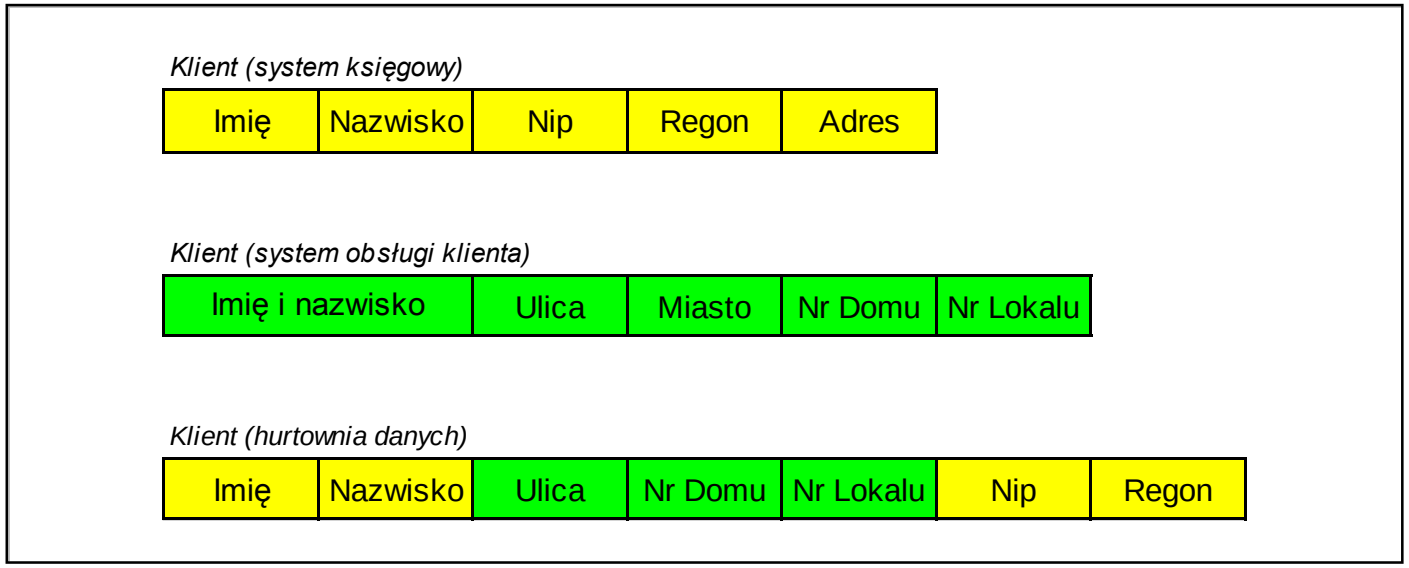
Identyfikacja źródeł danych ma wpływ na ogół późniejszych prac nad procesem ETL. Na jej etapie podejmowane są decyzje dotyczące kształtu systemu zasilania i organizacji prac związanych z jego budową.

Zespół ETL powinien zidentyfikować i udokumentować każdy system źródłowy wykorzystywany do zasilenia hurtowni danych.

Dokumentacja powstała w procesie analizy powinna być stale uaktualniania tak, aby w dowolnym momencie tworzenia procesu zasilania dawać zgodny z rzeczywistością obraz źródeł danych.

Dane z systemów operacyjnych

- Zazwyczaj dane dotyczące jednego tematu są gromadzone w wielu systemach, należy więc:
  - określić i udokumentować gdzie występują dane
  - oszacować i udokumentować jakość poszczególnych źródeł danych
  - określić możliwości dostępu do poszczególnych źródeł danych
  - określić skąd weźmiemy dane do hurtowni



Rodowód i profilowanie

Systemy źródłowe mogą posiadać dane zduplikowane, niejednokrotnie kopiowane, przenoszone, czyszczone i transformowane. Z tego względu do ich analizy należy podejść z ostrożnością

Przy identyfikacji powyższych problemów można wykorzystać narzędzia do ustalania rodowodu (pochodzenie, transformacje i reguły biznesowe, jakim podlegał element danych) oraz profilowania danych

Column Statistics :

Column N...	#Distinct	%Distinct	#Duplicate	%Duplicate	AVG	MIN	MAX	INFERRE...	#Null
Id	158	100	0	0	N/A	1	99	99	0
Fullname	156	99	2	1	N/A	Al Sneller	Yvette Tho...	No domain...	0
Title	75	47	83	53	N/A	Account Ma...	Vice Presid...	No domain...	43

Systemy źródłowe

Przy analizie systemów źródłowych należy zwrócić uwagę na:

- systemy współpracujące z systemem źródłowym
- grupy użytkowników korzystających z systemu
- dostawcę bazy danych, na której oparty jest system
- serwer produkcyjny bazy danych wraz z systemem operacyjnym
- obciążenie systemu w godzinach roboczych i poza nimi
- rozmiar bazy danych
- liczba dziennych transakcji

Zawartość danych

Przykładowe anomalie zawartości danych

- **wartości NULL** – przy złączeniach powodują poważne straty w danych, jeśli znajdują się w kluczach obcych. W bazie relacyjnej NULL nie jest równe NULL. W przypadku pojawienia się takich wartości można użyć złączeń zewnętrznych (outer joins), ale najlepszą praktyką jest nadanie im znaczenia biznesowego i przypisanie konkretnej wartości np. „nie dotyczy”

- **daty w polach o typie innym niż typ daty** – daty pojawiają się w różnych formatach, przyjmując różne wartości przy zachowaniu tego samego znaczenia. Np. w polu tekstowym mogą pojawić się warianty:

  25-Maj-2006

  25:05:2006 15:47:12

  Maj 06

Fazy projektowania systemu zasilania

1. opracowanie architektury systemu zasilania
2. selekcja źródeł danych
3. strategia pobierania danych
4. opracowanie transformacji
5. opracowanie harmonogramów wykonania
6. określenie sytuacji wyjątkowych i reakcji na nie

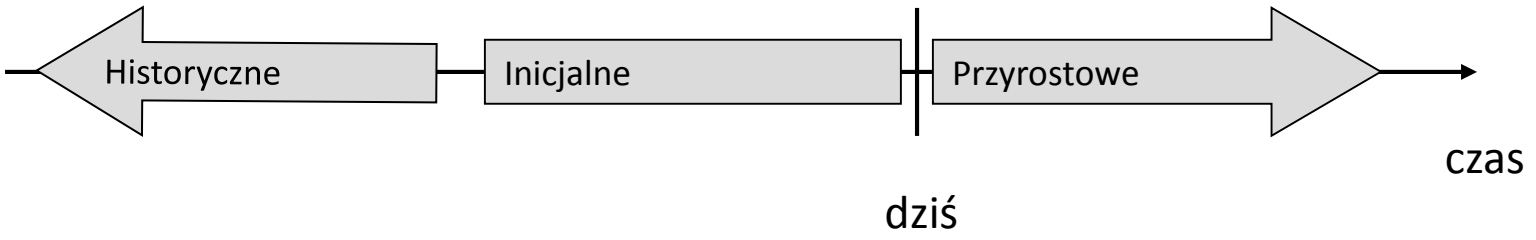
Ekstrakcja danych źródłowych

Inicjalna (migracja danych)

- pełen zakres danych
  - często na początku ograniczamy się do danych łatwo dostępnych (z aktualnego systemu operacyjnego), a dopiero później staramy się uzupełnić dane historyczne

Przyrostowa (cykliczna)

- potrzebujemy jedynie danych, które są nowe albo uległy zmianom



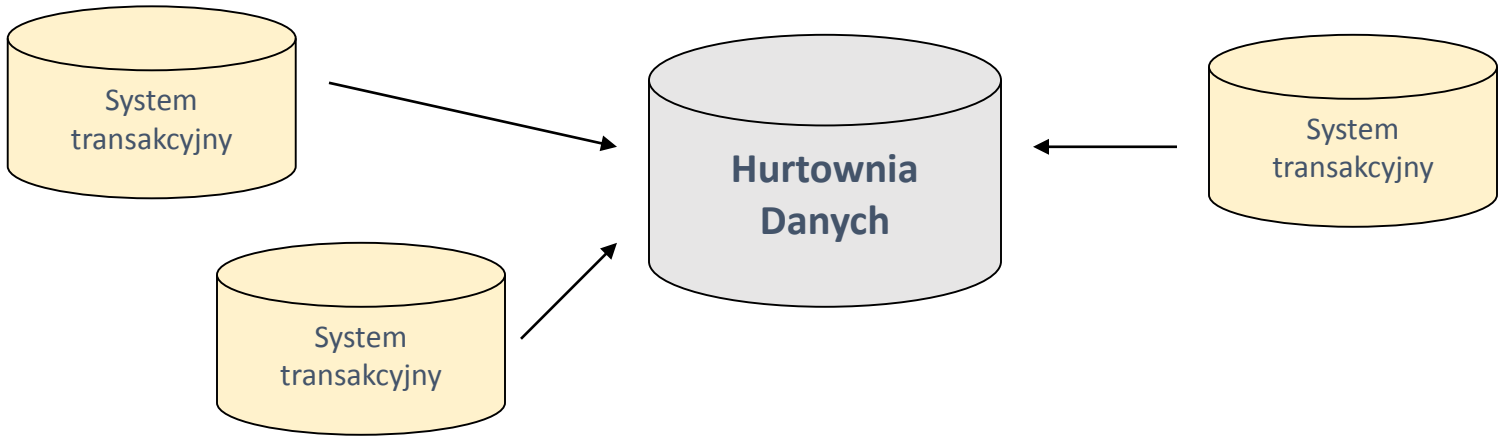
Migracja - Integracja i uspójnianie

Integracja danych

- zazwyczaj, dane dotyczące jednego zagadnienia, są gromadzone w wielu systemach źródłowych organizacji
- proces *integracji* danych polega na łączeniu podczas zasilania danych z wielu źródeł, tak aby w Hurtowni znalazły się wszystkie wymagane informacje, zorganizowane tematycznie.

Uspójnianie danych

- podczas integracji dane z różnych źródeł muszą zostać uspójnione (sprowadzenie do „wspólnego mianownika”)



Ekstrakcja przyrostowa i pełna

Pobranie całej zawartości jest dobre dla małych tabel

- dla dużych wolumenów danych jest często niewykonalna, ponieważ cykl ich zasilania trwa za długo
- zazwyczaj, tę strategię działania stosujemy dla słowników, prawie nigdy dla faktów

Ekstrakcja nowych transakcji jest zazwyczaj trudnym zadaniem, ale często jest jedynym akceptowalnym rozwiązaniem.

- Możliwe rozwiązania:
  - zastosowanie pola *„timestamp”* lub *sekwencji*
  - automatyczne rejestrowanie zmian
  - odczyt “logów” baz danych
  - …

Fazy projektowania systemu zasilania

1. opracowanie architektury systemu zasilania
2. selekcja źródeł danych
3. strategia pobierania danych
4. opracowanie transformacji
5. opracowanie harmonogramów wykonania
6. określenie sytuacji wyjątkowych i reakcji na nie

Techniki obróbki danych (1)

Przykłady technik podstawowych

- filtrowanie zbioru danych według określonych kryteriów
- agregacje – uporządkowanie zbioru danych w grupy i/lub wyliczenie podsumowań dla wyodrębnionych grup
- złączenia co najmniej dwóch zbiorów danych na zasadzie dopasowania elementów jednego zbioru do elementów drugiego zbioru według klucza
- unie co najmniej dwóch zbiorów – dodawanie zbiorów o tej samej strukturze
- sortowanie zbiorów danych
- generowanie sekwencji np. kluczy unikalnych
- wywołanie procedury składowanej
- funkcje rankingu

Techniki podstawowe mają odzwierciedlenie w komendach SQL jak i podstawowych obiektach narzędzi ETL

Techniki obróbki danych (2)

Narzędzia ETL oferują funkcjonalność, która nie ma odzwierciedlenia w pojedynczych komendach SQL (można je osiągnąć większym nakładem pracy), np.:

- podgląd zbiorów danych – możliwość odwołania się do zbioru w czasie wykonania mapowania np. w celu sprawdzenia czy istnieją w zbiorze określone wartości (również w wersji dynamicznej tzn. wraz z wartościami dołączanymi do zbioru w trakcie wykonania)
- kierowanie strumieniem danych – możliwość warunkowego podziału zbioru danych na podgrupy i skierowanie każdej z nich w dowolne miejsce
- kontrola transakcji – daje możliwość podziału zbioru na podgrupy, które można zapisać lub cofnąć w różnych transakcjach

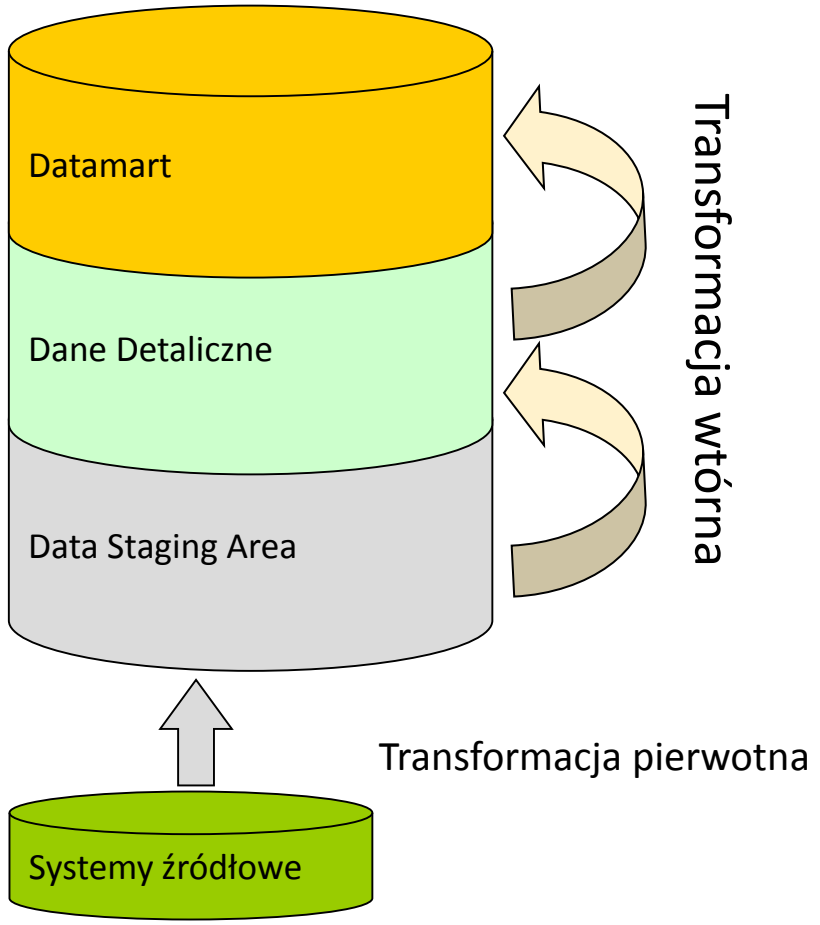
Typy transformacji

Pierwotna transformacja danych:

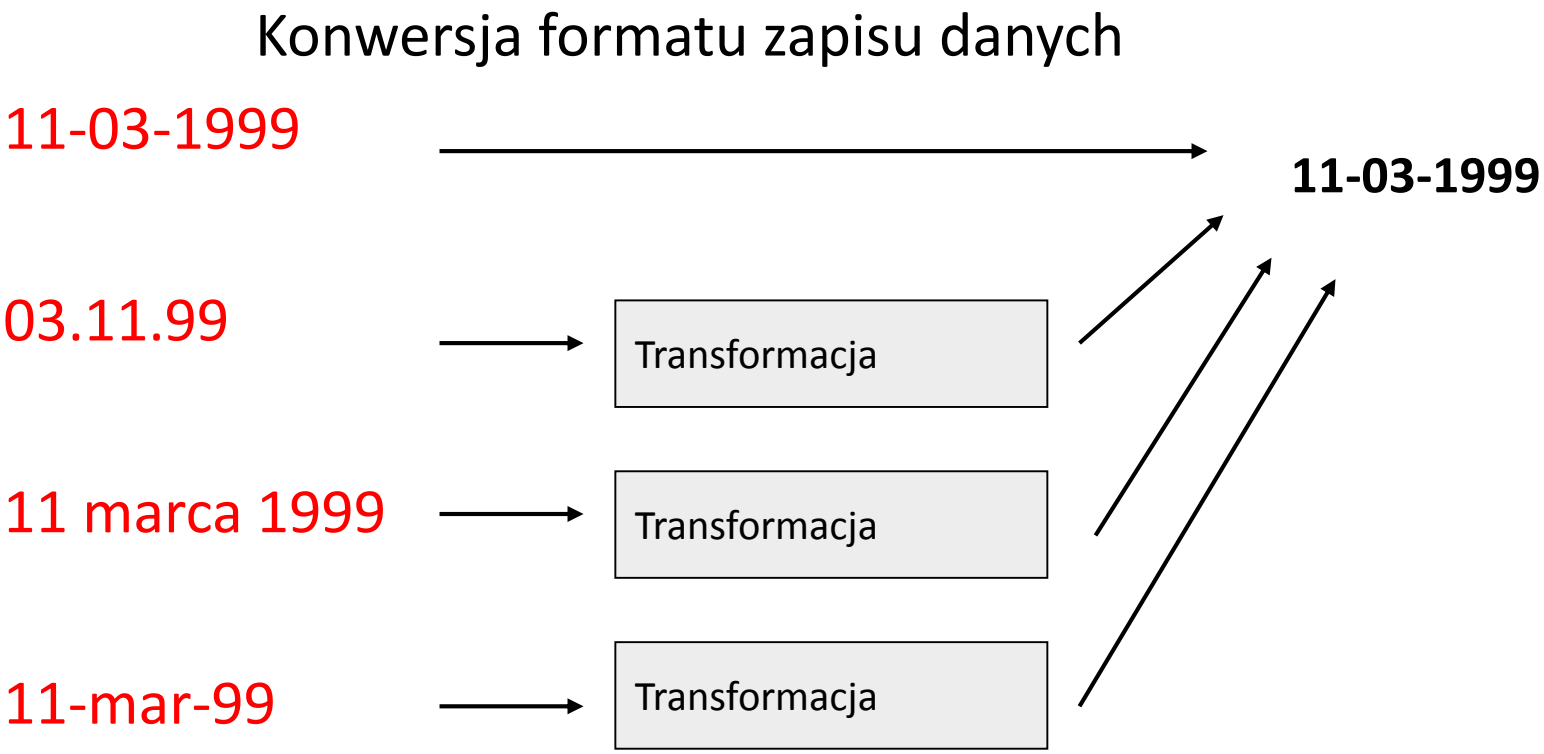
- pobieranie danych z systemów źródłowych (odczyt z systemów)
- walidacja
- integrowanie
- uspójniane
- czyszczenie
- przenoszenie do obszaru Hurtowni Danych (zapis do struktur)

Wtórna transformacja danych:

- agregowanie
- przeliczanie
- przenoszenie do hurtowni tematycznych



Przykładowe transformacje danych – ujednolicanie formatu danych

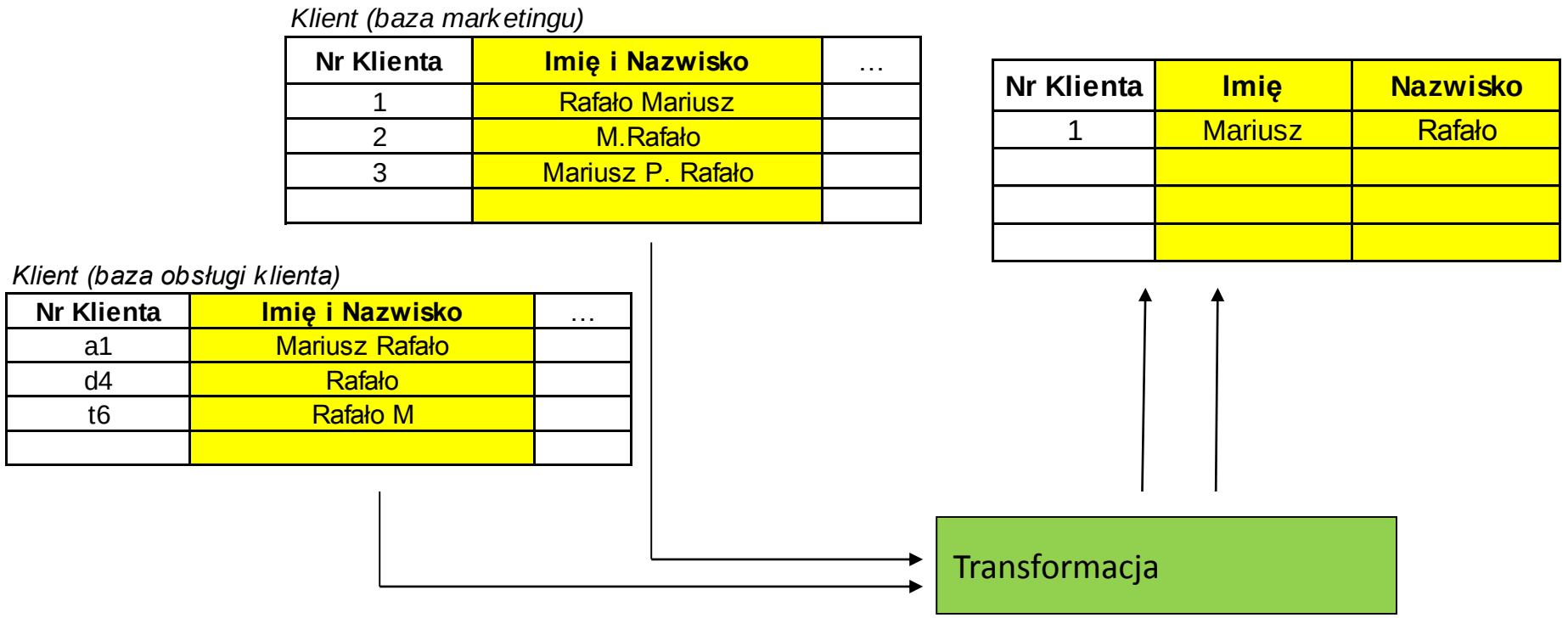


Przykładowe transformacje danych - konwersja

Konwersja jednostek

Przykładowe transformacje danych - duplikaty

Usuwanie duplikatów danych



Fazy projektowania systemu zasilania

1. opracowanie architektury systemu zasilania
2. selekcja źródeł danych
3. strategia pobierania danych
4. opracowanie transformacji
5. opracowanie harmonogramów wykonania
6. określenie sytuacji wyjątkowych i reakcji na nie

Zasilanie cykliczne vs zasilanie w czasie rzeczywistym

Zasilanie z dużym opóźnieniem (np. 24h) zazwyczaj jest realizowane wsadowo w trybie „pull” (na żądanie)

- narzędzia ETL cyklicznie odpytują systemy źródłowe o zmiany w danych
- tak działają tradycyjne rozwiązania ETL

Zasilanie w czasie rzeczywistym (real-time) polega na ładowaniu danych operacyjnych na bieżąco (on-line) w trybie „push” (zdarzeniowo)

- zmiany w systemach źródłowych są automatycznie propagowane do repozytorium skonsolidowanego
- w tym kierunku ewoluują narzędzia ETL

Kiedy stosować podejście zdarzeniowe

Podejście zdarzeniowe

- umożliwia działanie bez opóźnień (lub prawie bez opóźnień)
  - Real Time, Near Real Time, Right Time
- zazwyczaj mniej się nadaje do migrowania dużych wolumenów danych
  - najczęściej jest stosowane jako uzupełnienie tradycyjnego zasilania
  - dla pewnych klas danych
  - w połączeniu z klasycznym zasilaniem, np.:
    - Zdarzenia zbierane przez 30 minut i zapisywane do pliku
    - Następnie plik ładowany z wykorzystaniem ETL

Znajduje szczególne zastosowanie w operacyjnych zastosowaniach BI, gdzie wymagane jest działanie na danych aktualnych

Wsparcie Real Time

Przy zasilaniu w czasie rzeczywistym wykorzystywane są:

- platformy integracyjne (EAI)
- rozwiązania Change Data Capture (CDC) umożliwiające detekcję i propagację zmian danych
- platformy przetwarzania strumieniowego (np. Apache Kafka, Apache Flink)

Dla źródeł relacyjnych najczęściej stosuje się:

- wykorzystanie logu transakcji
- odpytywanie bazy danych
- wyzwalacze (trigery)

Fazy projektowania systemu zasilania

1. opracowanie architektury systemu zasilania
2. selekcja źródeł danych
3. strategia pobierania danych
4. opracowanie transformacji
5. opracowanie harmonogramów wykonania
6. określenie sytuacji wyjątkowych i reakcji na nie

Obsługa sytuacji awaryjnych

Podczas projektowania systemu zasilania należy:

- zidentyfikować możliwe sytuacje awaryjne, np.:
  - błąd danych źródłowych
  - brak danych źródłowych
  - błąd dostępu do źródeł danych
  - awaria sprzętowa
  - ...
- określić procedury administracyjne związane z ich obsługą

Cechy narzędzia ETL:

- automatyczna próba wznowienia sesji,
- wznawianie sesji (od miejsca zatrzymania)
- możliwość pracy w klastrze
  - przypisania wykonania sesji heterogenicznym węzłom w domenie (równolegle)
  - przeniesienie sesji na inny serwer

Dziękuję za uwagę