POLITECHNIKA GDAŃSKA

KATEDRA INŻYNIERII OPROGRAMOWANIA

Proces ETL

Krzysztof Goczyła
Teresa Zawadzka

Katedra Inżynierii Oprogramowania
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska
{kris, tegra}@eti.pg.gda.pl

POLITECHNIKA GDAŃSKA

Proces ETL

Wydobywanie danych (*Extract*)

- identyfikacja danych potrzebnych do analiz
- identyfikacja źródeł tych danych
- opracowanie procedur wydobywania danych

Przekształcanie danych (*Transform*)

- opracowanie odwzorowań pomiędzy danymi źródłowymi a docelowymi (formaty danych, jednostki miar, skalowanie, …)
- opracowanie zasad czyszczenia danych (dane odstające, brakujące, …)

Ładowanie danych (*Load*)

- przygotowanie pamięci (obszarów) na dane (*staging area*)
- opracowanie procedur ładowania
- ładowanie danych do tablic wymiarów
- ładowanie danych do tablic faktów
- tworzenie kostek

POLITECHNIKA GDAŃSKA

Data Profiling

Statystyczna analiza i ocena danych po kątem ich dalszego wykorzystania w projekcie.

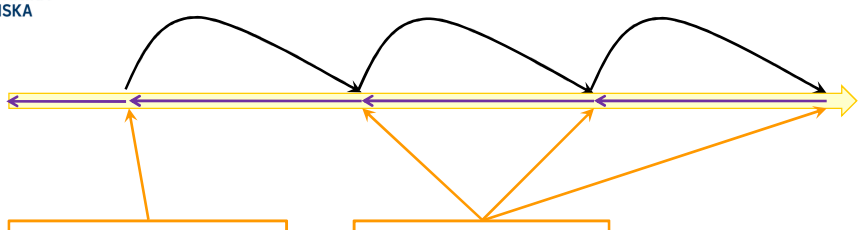
1. Dane, z których źródeł danych zostaną załadowane do hurtowni danych.
2. Identyfikacja jak największej liczby możliwych problemów z danymi.
3. Identyfikacja **metadanych**.
4. …

Dane o danych:
- *Kto jest autorem danych?*
- *Jaka jest struktura danych?*
- *Kiedy dane powstały?*
- *…*

- 3 -

POLITECHNIKA GDAŃSKA

Change Data Capture System



Pierwsze ładowanie:

Wszystkie dane historyczne zostają załadowane

Ładowanie co określony przedział czasowy.

Załaduj tylko te dane, które zmieniły się od ostatniego ładowania.

- 4 -

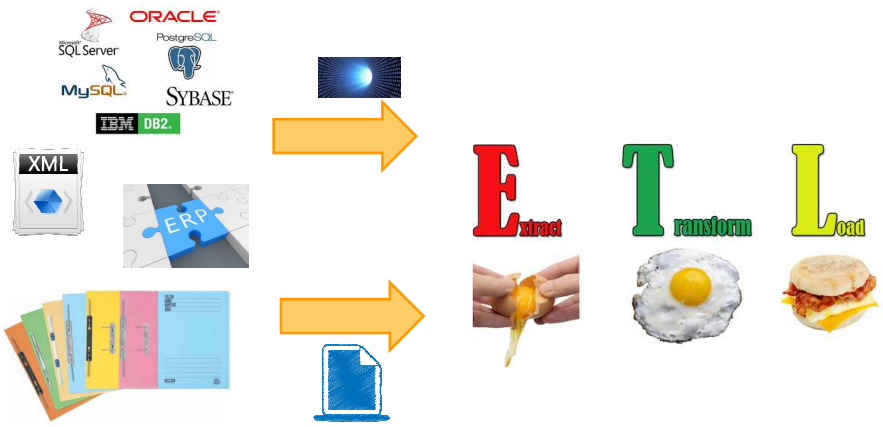
POLITECHNIKA GDAŃSKA

Change Data Capture System – podstawowe zadania

1. Wyizolowanie źródeł danych, w których nastąpiła zmiana.
2. Uchwycenie wszystkich zmian w danych, również tych wykonanych niestandardowymi interfejsami.
3. Oznacznikowanie danych w celu wyodrębnienia zmian związanych z poprawą błędów, a nie rzeczywistymi zmianami danych.
4. …

POLITECHNIKA GDAŃSKA

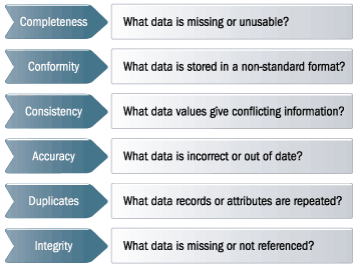
Extract System



POLITECHNIKA GDAŃSKA

Data Cleansing System

1. Wczesna diagnoza i ocena jakości danych.
2. Określenie wymagań na systemy źródłowe i zdefiniowanie procedury umożliwiających dostarczenie lepszych jakościowo danych.
3. Dostarczenie opisu błędów w danych możliwych do napotkania przez proces ETL.
4. Zaprojektowanie rozwiązania umożliwiającego wychwycenia wszystkich błędów związanych z jakością danych i precyzyjne monitorowanie metryk jakościowych w czasie.



POLITECHNIKA GDAŃSKA

Quality Screens (1)

Column screens — *Testowanie danych w ramach jednej kolumny, np.: czy występuje niedozwolona wartość null, czy wartość jest poza zakresem, itd.*

Structure screens — *Testowanie zależności pomiędzy dwoma lub więcej kolumnami, np.: zależności klucz główny/klucz obcy, itd.*

Business rules screens — *Złożone testowanie zależności biznesowych*

POLITECHNIKA GDAŃSKA

Quality Screens (2)

Quality screens

- *Zatrzymanie procesu*
- *Przesłanie podejrzanych danych do oddzielnego pliku w celu ich późniejszego przetworzenia*
- *Oznacznikowanie danych i przesłanie ich do następnego etapu procesu*

- 9 -

POLITECHNIKA GDAŃSKA

Generowanie elementów wymiarów

Elementy wymiarów są wartościami opisowymi!

Cena	Przedział cenowy
Data urodzenia	Kategoria wiekowa

- 10 -

POLITECHNIKA GDAŃSKA

Slowly Changing Dimension Manager

Id pacjenta	PESEL	nazwisko i imię	Płeć	…	Wiek	Data wstawienia	Data aktualizacji
4	70121201045	Zawadzka Teresa	Kobieta	…	Pomiędzy 25 a 35 rokiem	15.11.2010	null

Pacjent
Id_pacjenta
PESEL (KB)
nazwisko i imię
płeć
czy_ubezpieczony
zawód
wiek
województwo
miasto
data_wstawienia
data_aktualizacji

- 11 -

POLITECHNIKA GDAŃSKA

Slowly Changing Dimension Manager

Id pacjenta	PESEL	nazwisko i imię	Płeć	…	Wiek	Data wstawienia	Data aktualizacji
4	70121201045	Zawadzka Teresa	Kobieta	…	Pomiędzy 25 a 35 rokiem	15.11.2010	**12.09.2015**

Id pacjenta	PESEL	nazwisko i imię	Płeć	…	Wiek	Data wstawienia	Data aktualizacji
89	70121201045	Zawadzka Teresa	Kobieta	…	**Pomiędzy 35 a 45 rokiem**	**12.09.2015**	null

Ulec zmianie może więcej niż jedna wartość atrybutu!

- 12 -

POLITECHNIKA GDAŃSKA

Surogate Key Generator

1. Generowanie kluczy surogatowych podczas procesu ETL.
2. Generowanie kluczy surogatowych w bazie danych.

POLITECHNIKA GDAŃSKA

Wymiary

Date/Time Dimension

1. Nie mają typowego źródła danych.
2. Są definiowane i ładowane przed uruchomieniem właściwego procesu ETL.

Wymiar Inne

1. Przy stałej, znanej i nie za dużej liczbie krotek może być ładowany przed uruchomieniem właściwego procesu ETL.
2. W przeciwnym przypadku jest generowany w trakcie procesu ETL.

POLITECHNIKA GDAŃSKA

Fakty (1)

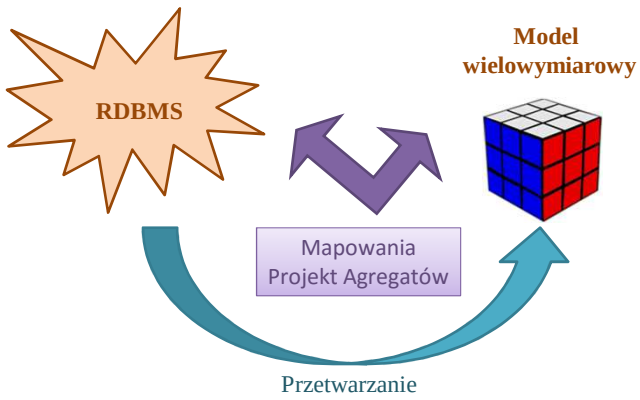
Zazwyczaj ładowane są po tabelach wymiarów.

W celu wpisania faktu do tabeli należy pobrać klucze surogatowe z tabel wymiarów, a następnie wyliczyć wartości miar.

- 15 -

POLITECHNIKA GDAŃSKA

Przetwarzanie kostki



- 16 -

POLITECHNIKA GDAŃSKA

Co każdy student wiedzieć powinien…

Podstawowe zadania systemu ETL