Technologie Zasilania i Odświeżania Hurtowni Danych

laboratorium

Krzysztof Jankiewicz
Politechnika Poznańska, Instytut Informatyki

Część 5

DATA WAREHOUSE STAGING AREA
TECHNIKI CHANGE DATA CAPTURE – PODSUMOWANIE

Data Warehouse Staging Area
co to jest?

- To miejsce składowania danych oraz zbiór procesów wykorzystywanych podczas przetwarzania ETL mających miejsce pomiędzy:
  - źródłami danych a hurtownią danych
  - hurtownią danych a data martami

Data Warehouse Staging Area
jakie ma cechy?

- Dane przechowywane w DSA:
  - mogą mieć charakter tymczasowy – po poprawnym załadowaniu danych mogą być kasowane,
  - mogą służyć jako słowniki – które np. są wykorzystywane przez procesy ETL,
  - mogą zawierać poprzednie obrazy danych źródłowych – np. w celu porównania ich z obecną ich postacią,
  - mogą zawierać dane pośrednie uzyskiwane podczas procesów ETL

Data Warehouse Staging Area
do czego może być wykorzystywany?

- Do przechowywania danych z różnych źródeł, które następnie będą **wielokrotnie wykorzystywane** w ramach procesów ETL.
- Do przechowywania danych wydobytych z systemów źródłowych w sposób szybki i prosty, aby następujące po tym procesy ETL **nie angażowały** systemów **źródłowych**.
- Do **wyszukania zmian** pomiędzy systemem źródłowym w obecnej i przeszłej postaci.
- Do wyszukania zmian w bieżącej postaci hurtowni danych i data martów.
- Do **transformacji** danych realizowanych **wieloetapowo** ze składowaniem pośrednich efektów przetwarzania.
- Do wyznaczania **agregatów**.
- Do wytworzenia i składowania danych w postaci docelowej dla hurtowni danych (w szczególności tabel faktów) w celu ich wydajnego ładowania.

Change Data Capture

- To proces wykrywania zmian w źródłach danych
- Podział CDC:
  - inwazyjne
    - oparte na danych źródła danych (ang. *Source-Based CDC*)
    - oparte na wyzwalaczach źródła danych (ang. *Trigger-Based CDC*)
    - oparte na porównaniu poprzedniego i bieżącego obrazu źródła danych (ang. *Snapshot-Based CDC*)
  - nieinwazyjne
    - oparte na dziennikach (np. plikach dzienników powtórzeń) (ang. *Log-Based CDC*)

Snapshot-Based CDC

- Porównanie dwóch wersji danych źródłowych – poprzedniej (przechowywanej zazwyczaj w OSA) oraz bieżącej
- Idea porównania polega na wyznaczeniu zmian w źródłowych danych
  - wstawionych – (`SELECT * FROM DANE_BIEZACE WHERE ID NOT IN (SELECT ID FROM DANE_POPRZEDNIE)`)
  - usuniętych – (`SELECT * FROM DANE_POPRZEDNIE WHERE ID NOT IN (SELECT ID FROM DANE_BIEZACE)`)
  - zmienionych – (`SELECT * FROM DANE_BIEZACE B JOIN DANE_POPRZEDNIE P ON (B.ID=P.ID) WHERE B.C1<>P.C1 OR B.C2<>P.C2 OR …`)

Log-Based CDC

- Analiza dzienników generowanych podczas pracy bazy danych
- Najmniej inwazyjny sposób wykrywania zmian
- Często używany w rozwiązaniach komercyjnych, ale także niekomercyjnych
  - Oracle GoldenGate
  - Attunity Stream
  - Wisdomforce
  - mysqlbinlog

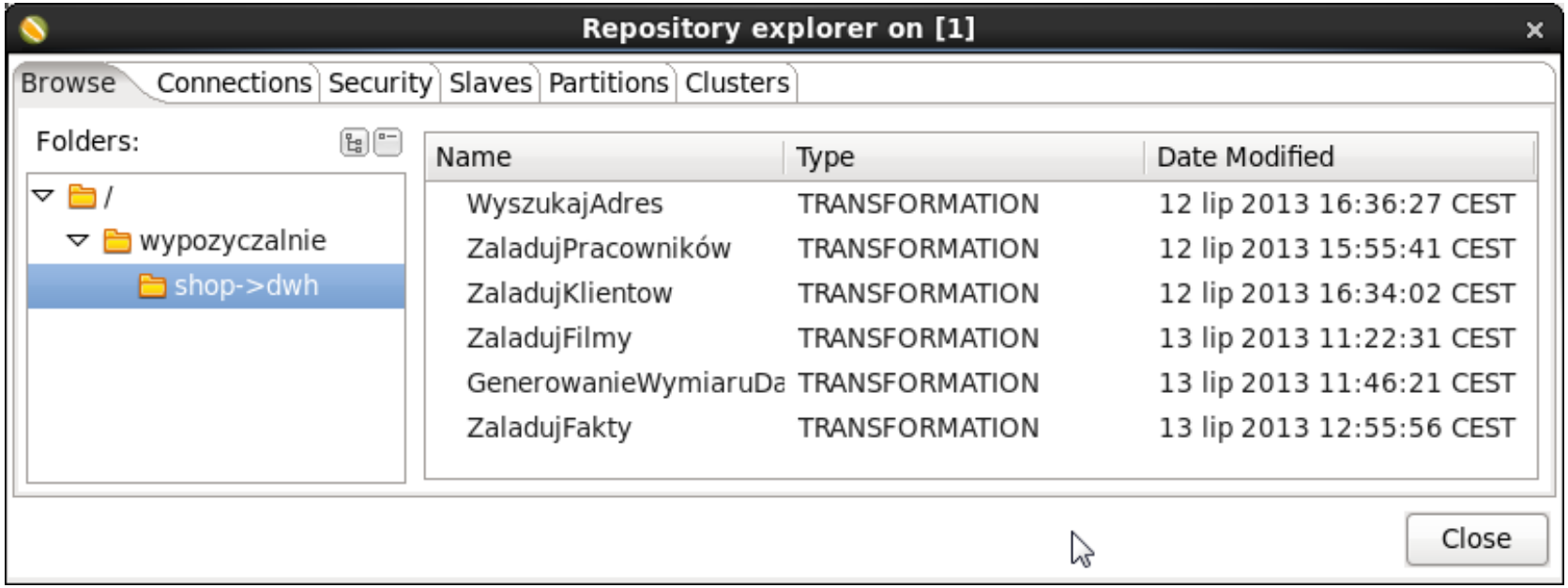
Cechy różnych typów CDC

Cechy	Source-Based	Trigger-Based	Snapshot-Based	Log-Based
Rozróżnienie pomiędzy wstawieniem a modyfikacją	Nie	Tak	Tak	Tak
Detekcja wielu (sekwencji) modyfikacji	Nie	Tak	Nie	Tak
Detekcja usunięcia	Nie	Tak	Tak	Tak
Bezinwazyjne	Nie	Nie	Nie	Tak
Wsparcie dla systemów czasu rzeczywistego	Nie	Tak	Nie	Tak
Wymagane zaangażowanie administratora bazy danych	Nie	Tak	Nie	Tak
Niezależne od bazy danych	Tak	Nie	Tak	Nie

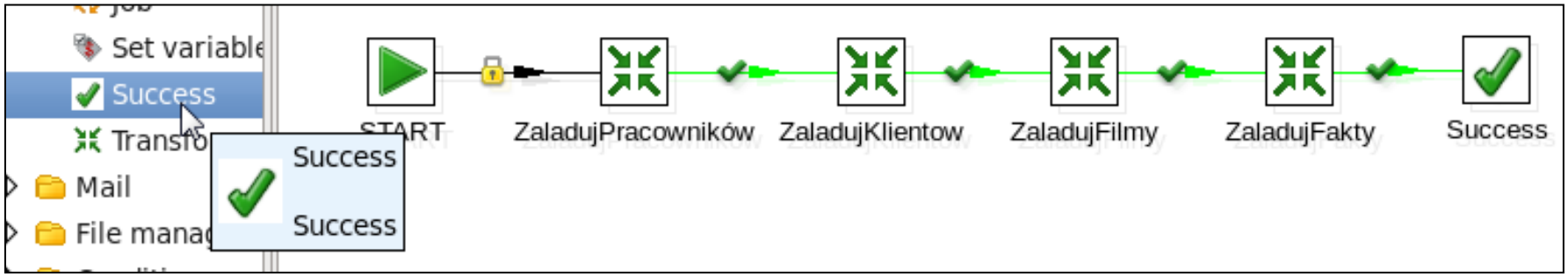
Omówienie ćwiczeń

- Zadania
  - Analiza utworzonych transformacji
  - Tworzenie zadań (ang. *jobs*)
  - Obserwacja efektów zadania
- Wykorzystanie sformatowanych plików jako źródeł danych
  - Analiza zawartości pliku CSV
  - Modyfikacja tabeli wymiaru klienci
  - Analiza planowanego rozwiązania
  - Utworzenie poprzedniego obrazu danych w obszarze DSA
  - Utworzenie transformacji implementującej Snaphot-based CDC
  - Utworzenie transformacji aktualizujących zawartość hurtowni danych
  - Integracja opracowanych transformacji za pomocą zadania.

Analiza utworzonych transformacji



Tworzenie zadań (ang. *jobs*)



Obserwacja efektów zadania

```
SQL> select count(*) from wypozyczenia;

  COUNT(*)
----------
      3467

SQL>
```

```
[etl@localhost ~]$ cd labs
[etl@localhost labs]$ ./zasilanie2.sh
```

```
SQL> select count(*) from wypozyczenia;

  COUNT(*)
----------
      3467

SQL> r
  1* select count(*) from wypozyczenia

  COUNT(*)
----------
     10176
```

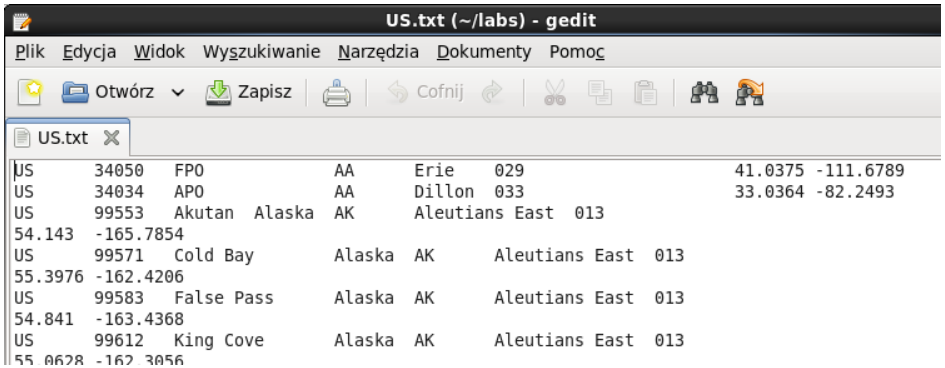
esults

History | Logging | Job metrics

Job / Job Entry	Comment	Result	Reason
OdswiezDWH			
Job: OdswiezDWH	Start of job execution		start
START	Start of job execution		start
START	Job execution finished	Success	
ZaladujPracowników	Start of job execution		Followed uncondition
ZaladujPracowników	Job execution finished	Success	
ZaladujKlientow	Start of job execution		Followed link after su
ZaladujKlientow	Job execution finished	Success	
ZaladujFilmy	Start of job execution		Followed link after su
ZaladujFilmy	Job execution finished	Success	
ZaladujFakty	Start of job execution		Followed link after su
ZaladujFakty	Job execution finished	Success	
Success	Start of job execution		Followed link after su
Success	Job execution finished	Success	
Job: OdswiezDWH	Job execution finished	Success	finished

Analiza zawartości pliku CSV

- country code – kod kraju iso
- postal code – kod pocztowy
- place name – nazwa miejscowości
- admin name1 – stan
- admin code1 – kod stanu
- admin name2 – okręg (prowincja)
- admin code2 – kod okręgu (prowincji)
- admin name3 – gmina (jednostka terytorialna)
- admin code3 – kod gminy (jednostki terytorialnej)
- latitude – szerokość geograficzna (wgs84)
- longitude – długość geograficzna (wgs84)
- accuracy – dokładność określenia długości i szerokości



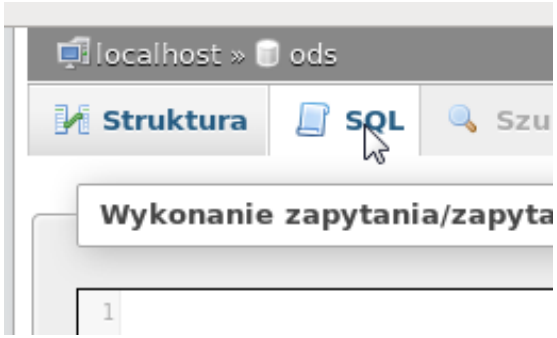
Modyfikacja tabeli wymiaru klienci

```
SQL> alter table klienci
  2  add KL_STAN VARCHAR2(100);

Table altered.

SQL> alter table klienci
  2  add KL_OKREG VARCHAR2(100);
```

Analiza planowanego rozwiązania

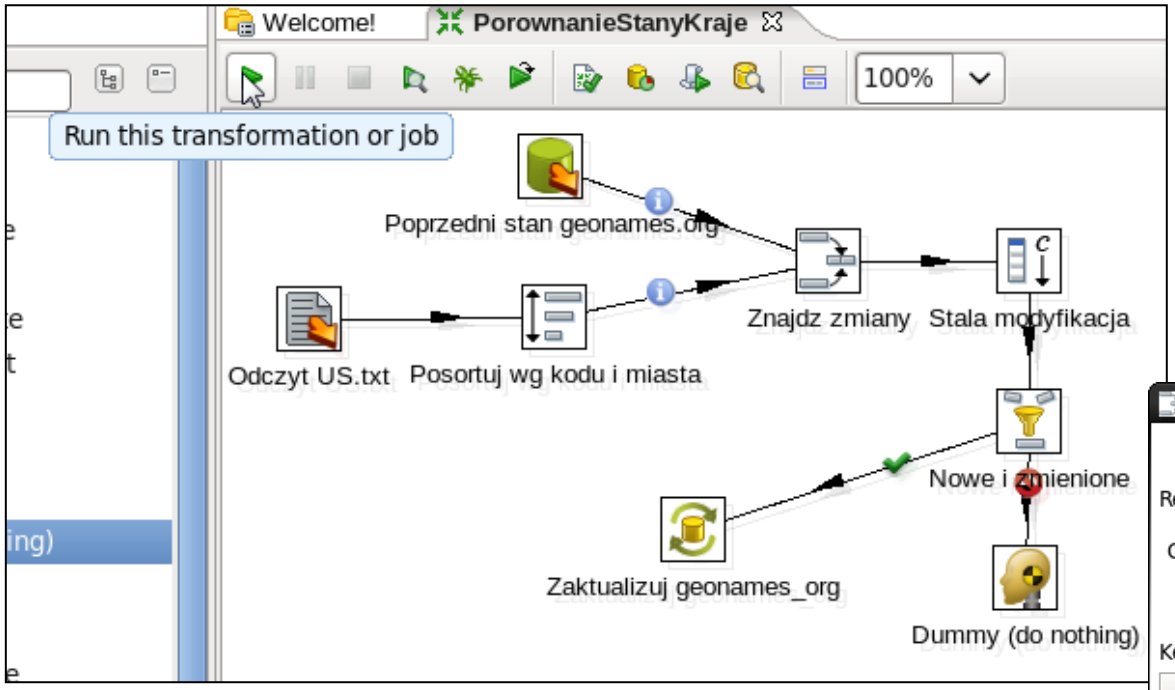
- Hurtownia danych już istnieje, dlatego nasza nowa transformacja powinna ten fakt uwzględniać i wprowadzić odpowiednie dane w jej zawartości
- Zawartość pliku CSV może się zmieniać w czasie, dlatego nie wystarczy aktualizować tylko danych, które nie posiadają określonego stanu i okręgu, ale także należy uwzględnić konieczność modyfikacji tych danych, które uległy zmianie (np. w wyniku zmian administracyjnych)
- Plik CSV nie posiada funkcjonalności, która pozwoliłaby wskazać zmienione składowe – musimy opracować metodę, która pozwoli je wykryć
- Uwzględniając przytoczone fakty opracujemy i zaimplementujemy w rzeczywistości dwie transformacje.
- Pierwsza z nich będzie przeznaczona do wykrywania zmian w źródle. Jej rezultatem będą odpowiednie znaczniki tych danych, które zostały zmienione pomiędzy kolejnymi aktualizacjami hurtowni danych.
- Druga z transformacji będzie aktualizowała dane wymiaru zarówno tych klientów,
  - którzy nie posiadają określonego stanu i okręgu (np. pojawiły się w ramach ostatniego odświeżenia hurtowni danych)
  - jak i tych klientów, dla których dane dotyczące stanu i okręgu uległy zmianie.

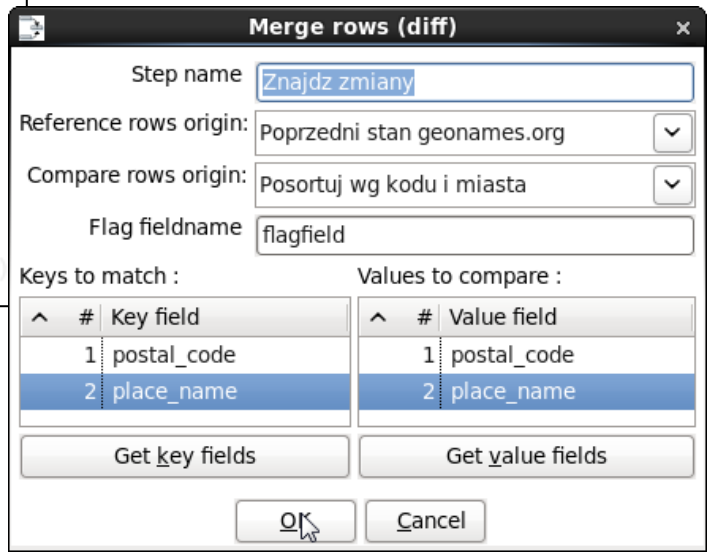
Utworzenie poprzedniego obrazu danych w obszarze DSA



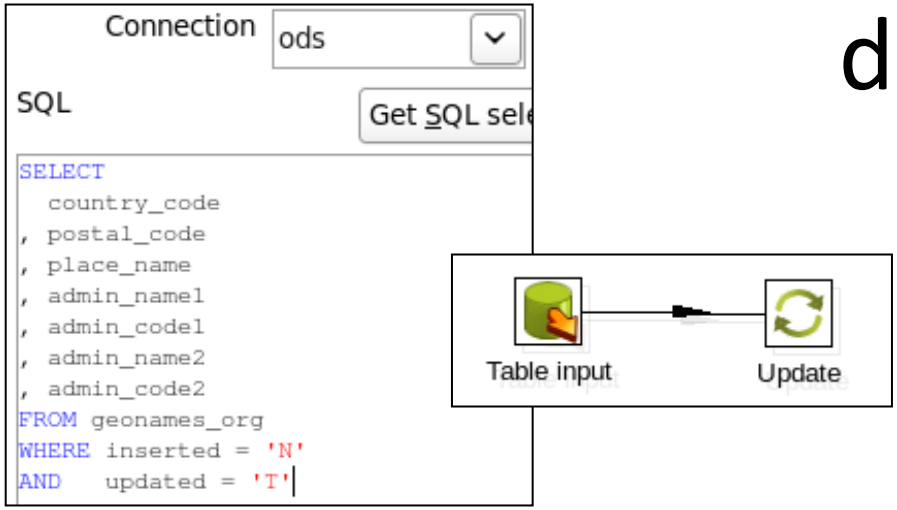

```
create table geonames_org (
  country_code varchar(2),
  postal_code  varchar(20),
  place_name   varchar(180),
  admin_name1  varchar(100),
  admin_code1  varchar(20),
  admin_name2  varchar(100),
  admin_code2  varchar(20),
  admin_name3  varchar(100),
  admin_code3  varchar(20),
  latitude     decimal(7,4),
  longitude    decimal(7,4),
  accuracy     decimal(7,4),
  inserted     char(1) default 'T',
  updated      char(1) );
```

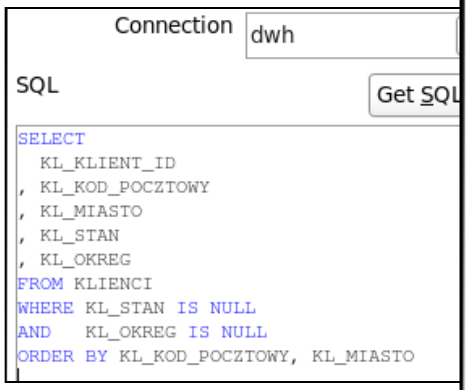
Utworzenie transformacji implementującej Snaphot-based CDC

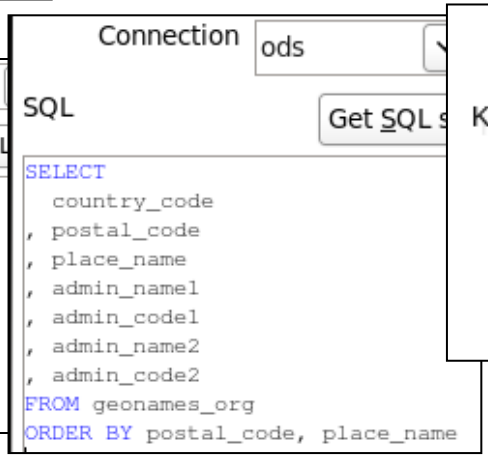


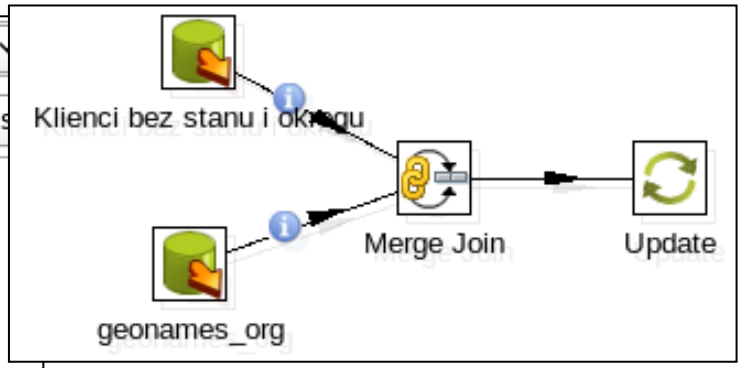


Utworzenie transformacji aktualizujących zawartość hurtowni danych

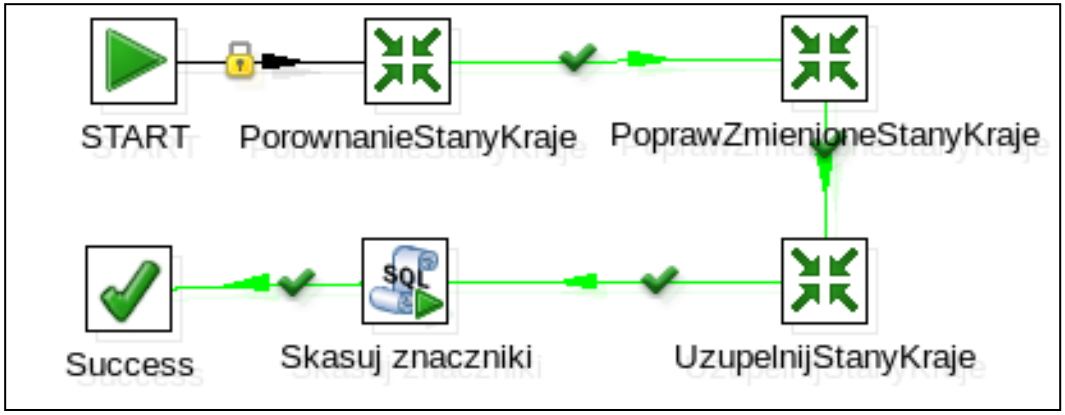








Integracja opracowanych transformacji za pomocą zadania



Część IV

ZEWNĘTRZNE ŹRÓDŁA DANYCH

Plan

- Wprowadzenie do formatu XML
- Strony WWW jako źródła danych
- Usługi sieciowe i ich typy
- Wprowadzenie do formatu JSON
- Usługi sieciowe jako źródła danych

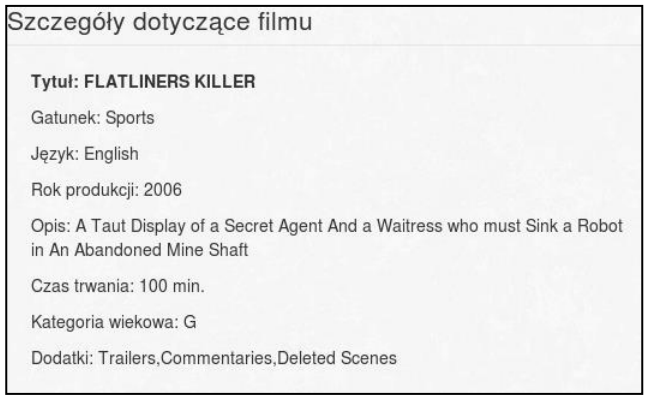
Wprowadzenie do XML

```xml
<?xml version="1.0" encoding="UTF-8"?>
<FILMY>
  <FILM id="1">
    <TYTUL>ACADEMY DINOSAUR</TYTUL>
    <OPIS>A Epic Drama of a Feminist And a Mad Scientist who must Battle a Teacher in The
Canadian Rockies</OPIS>
    <DATA_PRODUKCJI>2006</DATA_PRODUKCJI>
    <JEZYK>English</JEZYK>
    <DLUGOSC>86</DLUGOSC>
    <KATEGORIA_WIEKOWA>PG</KATEGORIA_WIEKOWA>
    <DODATKI>Deleted Scenes,Behind the Scenes</DODATKI>
    <GATUNEK>Documentary</GATUNEK>
  </FILM>
  <FILM id="2">
    <TYTUL>ADAPTATION HOLES</TYTUL>
    <OPIS>A Astounding Reflection of a Lumberjack
Baloon Factory</OPIS>
    <DATA_PRODUKCJI>2006</DATA_PRODUKCJI>
    <JEZYK>English</JEZYK>
    <DLUGOSC>50</DLUGOSC>
    <KATEGORIA_WIEKOWA>NC-17</KATEGORIA_WIEKOWA>
    <DODATKI>Trailers,Deleted Scenes</DODATKI>
    <GATUNEK>Documentary</GATUNEK>
  </FILM>
. . .
```

```xml
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="FILMY">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" ref="FILM"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="FILM">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="TYTUL"/>
        <xs:element ref="OPIS"/>
        <xs:element ref="DATA_PRODUKCJI"/>
        <xs:element ref="JEZYK"/>
        <xs:element ref="DLUGOSC"/>
        <xs:element ref="KATEGORIA_WIEKOWA"/>
        <xs:element ref="DODATKI"/>
        <xs:element ref="GATUNEK"/>
```

Strony WWW jako źródła danych

- Zewnętrzne źródło wartościowych danych
- Wiele różnorodnych standardów:
  - HTML
  - HTTP
  - XML
  - XSL
  - CSS



Usługi sieciowe i ich typy

- Powszechnie stosowana metoda integracji systemów informatycznych
- Dwa główne paradygmaty tworzenia usług sieciowych
  - Big-Web Services – oparte na protokole SOAP, złożone, wykorzystujące wiele różnorodnych specyfikacji i standardów
  - RESTful – oparte na protokole HTTP, proste, zorientowane na przetwarzanie zasobów, różnorodne formaty danych: XML, JSON, TXT

JSON (JavaScript Object Notation)

- Prosty format danych
  - XML to prawie 90 reguł gramatyki języka
  - JSON to tylko 15 reguł
- Zyskuje dużą popularność, np. powszechnie wykorzystywany w przypadku usług RESTful (na równi z XML)

Usługi sieciowe jako źródła danych

- Nie tylko funkcjonalność – do integracji aplikacji
- Ale także dane – do dystrybucji informacji (w szczególności dotyczy to usług RESTful)

{"row":[{"ACTOR_ID":23,"FIRST_NAME":"SANDRA","LAST_NAME":"KILMER"},{"ACTOR_ID":37,"FIRST_NAME":"VAL","LAST_NAME":

```
http://localhost:8181/apex/apex_rest.getReport?app=AKTORZY&
page=1&reportid=AKTORZY&output=JSON&parmvalues=YOUTH%20KICK
```

Omówienie ćwiczeń

- Dokument XML jako źródło danych
  - Analiza zawartości dokumentu XML
  - Dostosowanie definicji hurtowni danych – rozbudowa wymiaru filmy
  - Transformacja zasilająca hurtownię danych na podstawie dokumentu XML
- Usługa sieciowa RESTful jako źródło danych
  - Analiza funkcjonalności usługi
  - Dostosowanie definicji hurtowni danych – rozbudowa wymiaru filmy
  - Transformacja zasilająca hurtownie na podstawie wywołań usługi RESTful

Analiza zawartości dokumentu XML

```
<?xml version="1.0" encoding="UTF-8"?>
<FILMY>
  <FILM id="1">
    <TYTUL>ACADEMY DINOSAUR</TYTUL>
    <OPIS>A Epic Drama of a Feminist And a Mad Scientist who must Battle a Teacher in
The Canadian Rockies</OPIS>
    <DATA_PRODUKCJI>2006</DATA_PRODUKCJI>
    <JEZYK>English</JEZYK>
    <DLUGOSC>86</DLUGOSC>
    <KATEGORIA_WIEKOWA>PG</KATEGORIA_WIEKOWA>
    <DODATKI>Deleted Scenes,Behind the Scenes</DODATKI>
    <GATUNEK>Documentary</GATUNEK>
  </FILM>
  <FILM id="2">
    <TYTUL>ADAPTATION HOLES</TYTUL>
    <OPIS>A Astounding Reflection of a L
in A Baloon Factory</OPIS>
    <DATA_PRODUKCJI>2006</DATA_PRODUKCJI
    <JEZYK>English</JEZYK>
    <DLUGOSC>50</DLUGOSC>
    <KATEGORIA_WIEKOWA>NC-17</KATEGORIA_
    <DODATKI>Trailers,Deleted Scenes</DO
    <GATUNEK>Documentary</GATUNEK>
  </FILM>
. . .
```

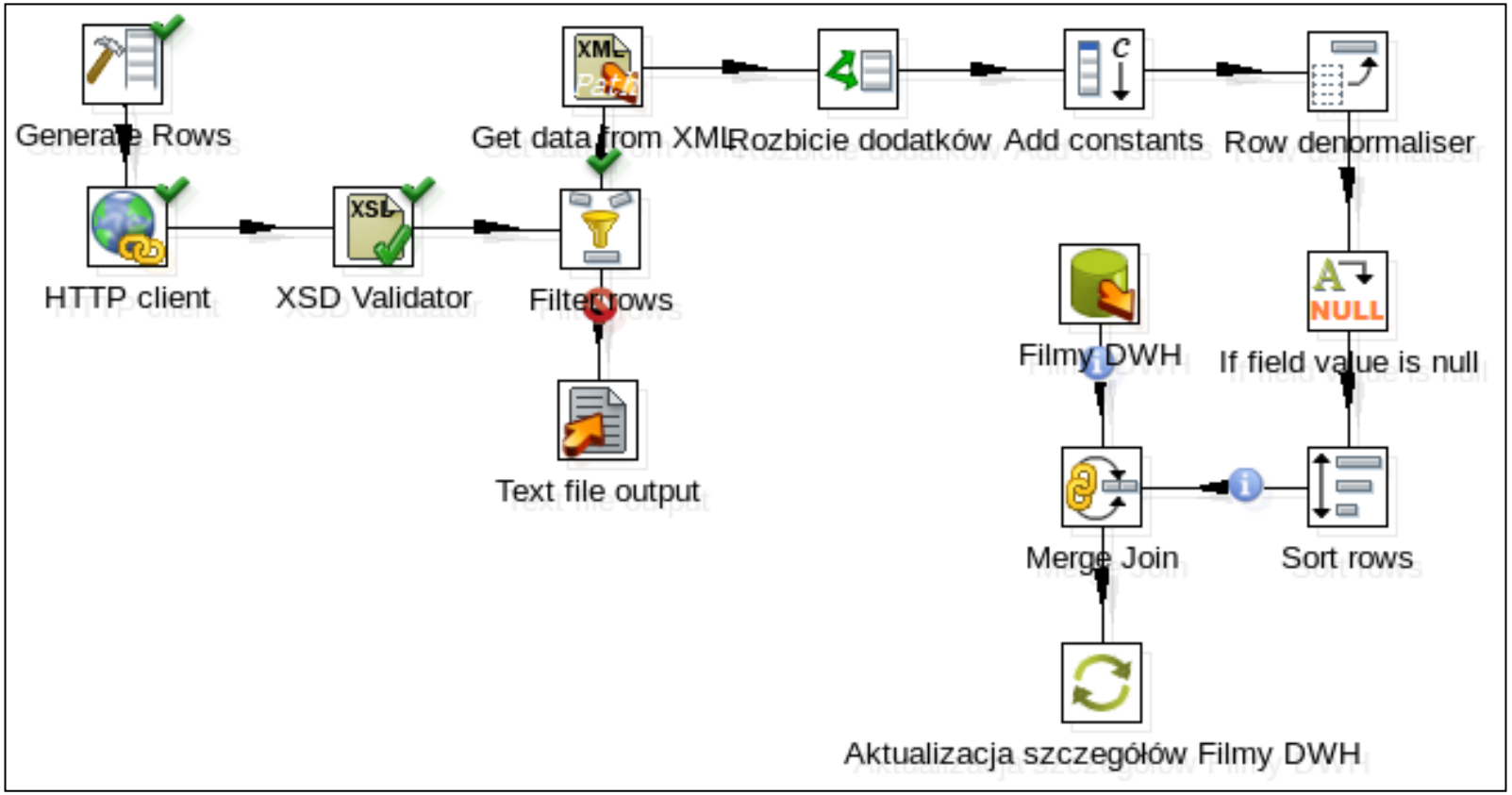
```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="FILMY">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" ref="FILM"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <xs:element name="FILM">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="TYTUL"/>
        <xs:element ref="OPIS"/>
        <xs:element ref="DATA_PRODUKCJI"/>
        <xs:element ref="JEZYK"/>
        <xs:element ref="DLUGOSC"/>
        <xs:element ref="KATEGORIA_WIEKOWA"/>
        <xs:element ref="DODATKI"/>
        <xs:element ref="GATUNEK"/>
```

Dostosowanie definicji hurtowni danych – rozbudowa wymiaru filmy

```
SQL> desc filmy
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------
 FI_FILM_ID                                NOT NULL NUMBER
 FI_OSTATNIA_MODYFIKACJA                   NOT NULL TIMESTAMP(6)
 FI_TYTUL                                  NOT NULL VARCHAR2(255)
 FI_ROK_WYDANIA                            NOT NULL NUMBER
```

```
alter table filmy
add ( FI_OPIS                 VARCHAR2(150),
      FI_JEZYK                VARCHAR2(10),
      FI_DLUGOSC              NUMBER,
      FI_KATEGORIA_WIEKOWA    VARCHAR2(5),
      FI_CZY_TRAILERY         CHAR(1),
      FI_CZY_KOMENTARZE       CHAR(1),
      FI_CZY_USUNIETE_SCENY   CHAR(1),
      FI_CZY_DODATKOWE_SCENY CHAR(1),
      FI_GATUNEK              VARCHAR2(25) );
```

Transformacja zasilająca hurtownię danych na podstawie dokumentu XML



Analiza funkcjonalności usługi RESTful

- Adres URL: **`http://localhost:8181/apex/apex_rest.getReport`**.
- Parametry:
  - **`app`** – nazwa aplikacji w ramach której stworzono raport udostępniający dane za pomocą usługi RESTful
  - **`page`** – identyfikator strony, na której wspomniany raport się znajduje
  - **`reportid`** – nazwa raportu udostępniającego dane
  - **`output`** – format danych wynikowych (XML lub JSON)
  - **`parmvalues`** – wartości parametrów wykorzystywanych przez raport do ograniczania wyników

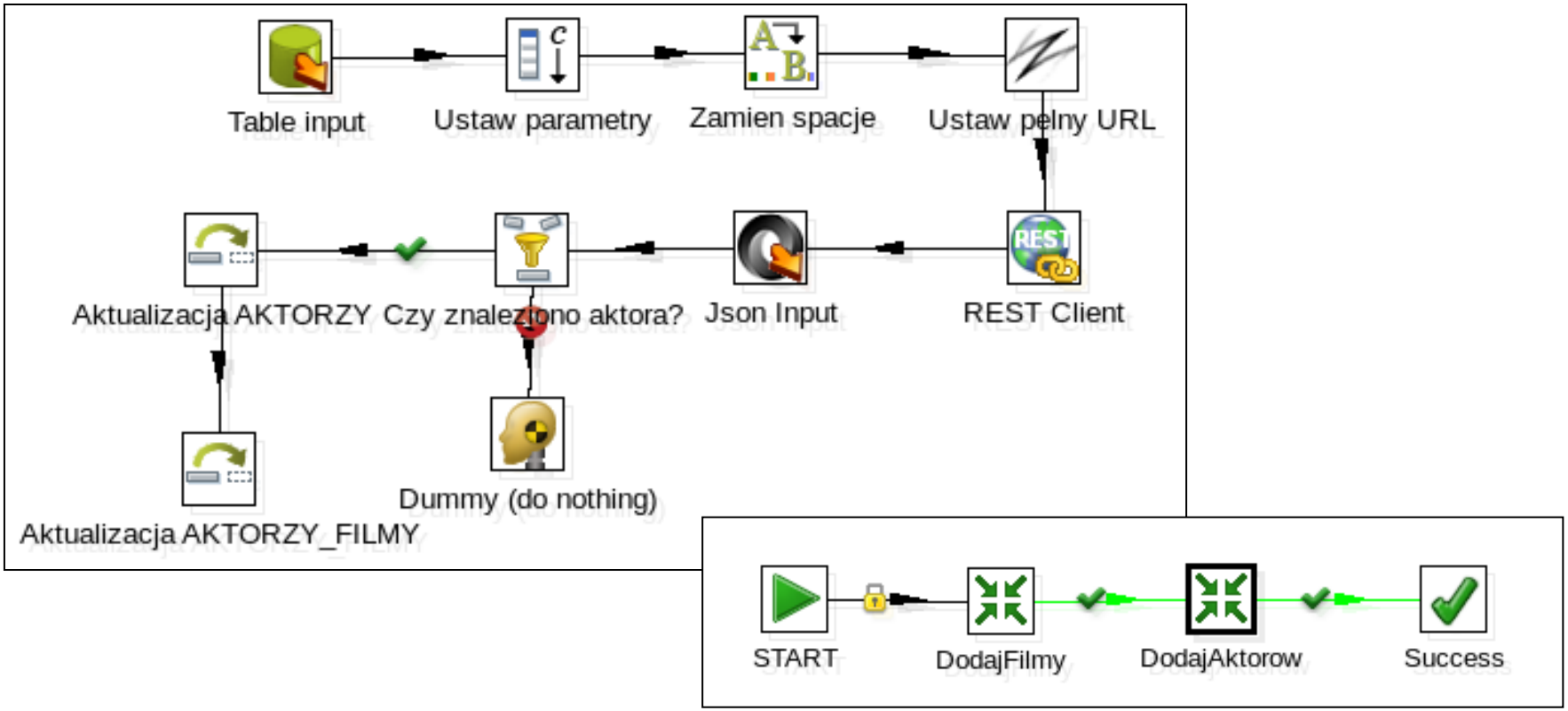
```
http://localhost:8181/apex/apex_rest.getReport?app=AKTORZY&
page=1&reportid=AKTORZY&output=JSON&parmvalues=YOUTH%20KICK
```

{"row":[{"ACTOR_ID":23,"FIRST_NAME":"SANDRA","LAST_NAME":"KILMER"},{"ACTOR_ID":37,"FIRST_NAME":"VAL","LAST_NAME":

Dostosowanie definicji hurtowni danych – rozbudowa wymiaru filmy

```
create table aktorzy (
  AK_AKTOR_ID NUMBER NOT NULL PRIMARY KEY,
  AK_NAZWISKO VARCHAR2(45) NOT NULL,
  AK_IMIE     VARCHAR2(45) NOT NULL);

create table aktorzy_filmy (
  AF_AKTOR_ID NUMBER NOT NULL
              REFERENCES aktorzy(AK_AKTOR_ID),
  AF_FILM_ID  NUMBER NOT NULL
              REFERENCES filmy(FI_FILM_ID),
  PRIMARY KEY (AF_AKTOR_ID, AF_FILM_ID) );
```

Transformacja zasilająca hurtownie na podstawie wywołań usługi RESTful



Część 7

PROFILOWANIE I CZYSZCZENIE DANYCH
SLOWLY CHANGING DIMENSION

Czyszczenie danych

- Czyszczenie danych to
  - bardzo złożone zagadnienie i
  - obejmujące wiele różnorodnych działań
  - mające wpływ na zawartość
    - nie tylko hurtowni danych, ale także
    - obszaru ODS czy też
    - źródeł danych
  - poprzedzane oceną jakości danych źródłowych – profilowaniem danych

Slowly Changing Dimension

- Z założenia hurtownia danych zawiera dane trwałe – nieulotne.
- Nieulotność wymusza uwzględnienie zmian w atrybutach wymiarów
- Zazwyczaj stosuje się następujące podejścia do tzw. wymiarów wolnozmiennych (ang. *Slowly Changing Dimension*)
  - Type 0 – brak jakichkolwiek działań
  - Type 1 – nadpisywanie
  - Type 2 – tworzenie nowych wersji danych
  - Type 3 – dodatkowe atrybuty przechowujące np. poprzednią wersję wartości wybranych atrybutów
  - Type 4 – przechowywanie poprzednich wersji danych w oddzielnych strukturach

Slowly Changing Dimension
type 1

Rows of step: Adres zmieniony (3 rows)

#	KL_KLIENT_ID	KL_KOD_POCZTOWY	KL_MIASTO	KL_ULICA	klient_id	miasto	adres	kod_pocztowy
1	285	26209	Snowshoe	1336 Benin City Drive	285	Los Angeles	98 Stara Zagora Boulevard	90001
2	342	65583	Waynesville	1293 Nam Dinh Way	342	Eckert	1192 Tongliao Street	81418
3	399	77850	Concord	734 Bchar Place	399	Vera	953 Hodeida Street	74082

poprzednie wartości atrybutów

obecne wartości atrybutów

przed aktualizacją

285	26209	Snowshoe	1336 Benin City Drive
342	65583	Waynesville	1293 Nam Dinh Way
399	77850	Concord	734 Bchar Place

po aktualizacji

285	90001	Los Angeles	28 Stara Zagora Boulevard
342	81418	Eckert	1192 Tongilao Street
399	74082	Vera	953 Hodeida Street

Slowly Changing Dimension

type 2

Rows of step: Adres zmieniony (3 rows)

#	KL_KLIENT_ID	KL_KOD_POCZTOWY	KL_MIASTO	KL_ULICA	klient_id	miasto	adres	kod_pocztowy
1	285	26209	Snowshoe	1336 Benin City Drive	285	Los Angeles	98 Stara Zagora Boulevard	90001
2	342	65583	Waynesville	1293 Nam Dinh Way	342	Eckert	1192 Tongliao Street	81418
3	399	77850	Concord	734 Bchar Place	399	Vera	953 Hodeida Street	74082

poprzednie wartości atrybutów

obecne wartości atrybutów

przed aktualizacją

KL_ID	KL_KOD	KL_MIASTO	KL_ULICA	KL_SID	KL_VER	KL_START	KL_STOP
285	26209	Snowshoe	1336 Benin City Drive	1	1	1900-01-01	2999-12-31
342	65583	Waynesville	1293 Nam Dinh Way	2	1	1900-01-01	2999-12-31
399	77850	Concord	734 Bchar Place	3	1	1900-01-01	2999-12-31

po aktualizacji

KL_ID	KL_KOD	KL_MIASTO	KL_ULICA	KL_SID	KL_VER	KL_START	KL_STOP
285	26209	Snowshoe	1336 Benin City Drive	1	1	1900-01-01	2013-07-23
342	65583	Waynesville	1293 Nam Dinh Way	2	1	1900-01-01	2013-07-23
399	77850	Concord	734 Bchar Place	3	1	1900-01-01	2013-07-23
285	90001	Los Angeles	28 Stara Zagora Boulevard	4	2	2013-07-23	2999-12-31
342	81418	Eckert	1192 Tongilao Street	5	2	2013-07-23	2999-12-31
399	74082	Vera	953 Hodeida Street	6	2	2013-07-23	2999-12-31

str. 41

Slowly Changing Dimension

type 3

Rows of step: Adres zmieniony (3 rows)

#	KL_KLIENT_ID	KL_KOD_POCZTOWY	KL_MIASTO	KL_ULICA	klient_id	miasto	adres	kod_pocztowy
1	285	26209	Snowshoe	1336 Benin City Drive	285	Los Angeles	98 Stara Zagora Boulevard	90001
2	342	65583	Waynesville	1293 Nam Dinh Way	342	Eckert	1192 Tongliao Street	81418
3	399	77850	Concord	734 Bchar Place	399	Vera	953 Hodeida Street	74082

poprzednie wartości atrybutów

obecne wartości atrybutów

przed aktualizacją

KL_ID	KL_KOD	KL_MIASTO	KL_ULICA	KL_KOD_OLD	KL_MIASTO_OLD	KL_ULICA_OLD
285	26209	Snowshoe	1336 Benin City Drive			
342	65583	Waynesville	1293 Nam Dinh Way			
399	77850	Concord	734 Bchar Place			

po aktualizacji

KL_ID	KL_KOD	KL_MIASTO	KL_ULICA	KL_KOD_OLD	KL_MIASTO_OLD	KL_ULICA_OLD
285	90001	Los Angeles	28 Stara Zagora Boulevard	26209	Snowshoe	1336 Benin City Drive
342	81418	Eckert	1192 Tongilao Street	65583	Waynesville	1293 Nam Dinh Way
399	74082	Vera	953 Hodeida Street	77850	Concord	734 Bchar Place

Slowly Changing Dimension

type 4

Rows of step: Adres zmieniony (3 rows)

#	KL_KLIENT_ID	KL_KOD_POCZTOWY	KL_MIASTO	KL_ULICA	klient_id	miasto	adres	kod_pocztowy
1	285	26209	Snowshoe	1336 Benin City Drive	285	Los Angeles	98 Stara Zagora Boulevard	90001
2	342	65583	Waynesville	1293 Nam Dinh Way	342	Eckert	1192 Tongliao Street	81418
3	399	77850	Concord	734 Bchar Place	399	Vera	953 Hodeida Street	74082

poprzednie wartości atrybutów

obecne wartości atrybutów

KLIENCI

przed aktualizacją

KL_ID	KL_IMIE	KL_NAZWISKO
285	Adam	Nowak
342	Piotr	Kowalski
399	Zofia	Zając

po aktualizacji

KL_ID	KL_IMIE	KL_NAZWISKO
285	Adam	Nowak
342	Piotr	Kowalski
399	Zofia	Zając

ADRESY_KLIENTOW

AD_SID	AD_KOD	AD_MIASTO	AD_ULICA
1	26209	Snowshoe	1336 Benin City Drive
2	65583	Waynesville	1293 Nam Dinh Way
3	77850	Concord	734 Bchar Place

AD_SID	AD_KOD	AD_MIASTO	AD_ULICA
1	26209	Snowshoe	1336 Benin City Drive
2	65583	Waynesville	1293 Nam Dinh Way
3	77850	Concord	734 Bchar Place
4	90001	Los Angeles	28 Stara Zagora Boulevard
5	81418	Eckert	1192 Tongilao Street
6	74082	Vera	953 Hodeida Street

str. 43

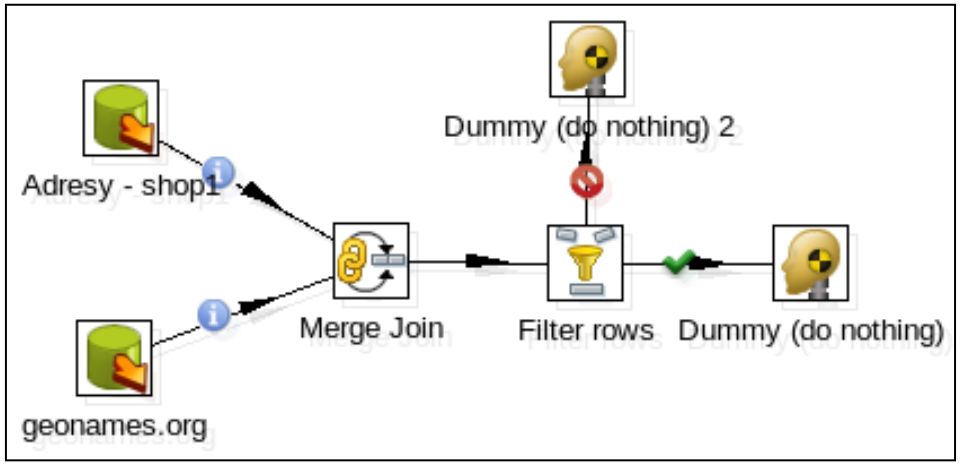
Slowly Changing Dimensions
wpływ na tabele faktów

- Różne CDC mają różny wpływ na tabelę faktów
  - Type 0 – brak zmian
  - Type 1 – brak zmian
  - Type 2 – tabela faktów posiada klucz obcy oparty na sztucznym kluczu głównym tabeli wymiarów (ang. *surrogate key*)
  - Type 3 – bez zmian
  - Type 4 – dodatkowa tabela wymiaru, ze sztucznym kluczem głównym, wymaga dodatkowego klucza obcego w tabeli faktów

Omówienie ćwiczeń

- Czyszczenie danych
  - Analiza problemu – profilowanie danych
  - Transformacja mająca na celu automatyczną naprawę danych adresowych w źródłach
  - Transformacja mająca na celu weryfikację poprawności adresów e-mail
  - Odświeżenie hurtowni danych – poprawa jakości danych w hurtowni
- *Slowly Changing Dimension*
  - Analiza zmian wartości atrybutów w wymiarach
  - Modyfikacja schematu hurtowni danych mająca na celu rozbudowanie wymiaru klient na potrzeby CDC type 2.
  - Modyfikacja transformacji dotyczącej wymiaru klient implementująca CDC type 2.
  - Modyfikacja transformacji dotyczącej faktów uwzględniająca zaimplementowaną CDC type 2 na wymiarze klient.

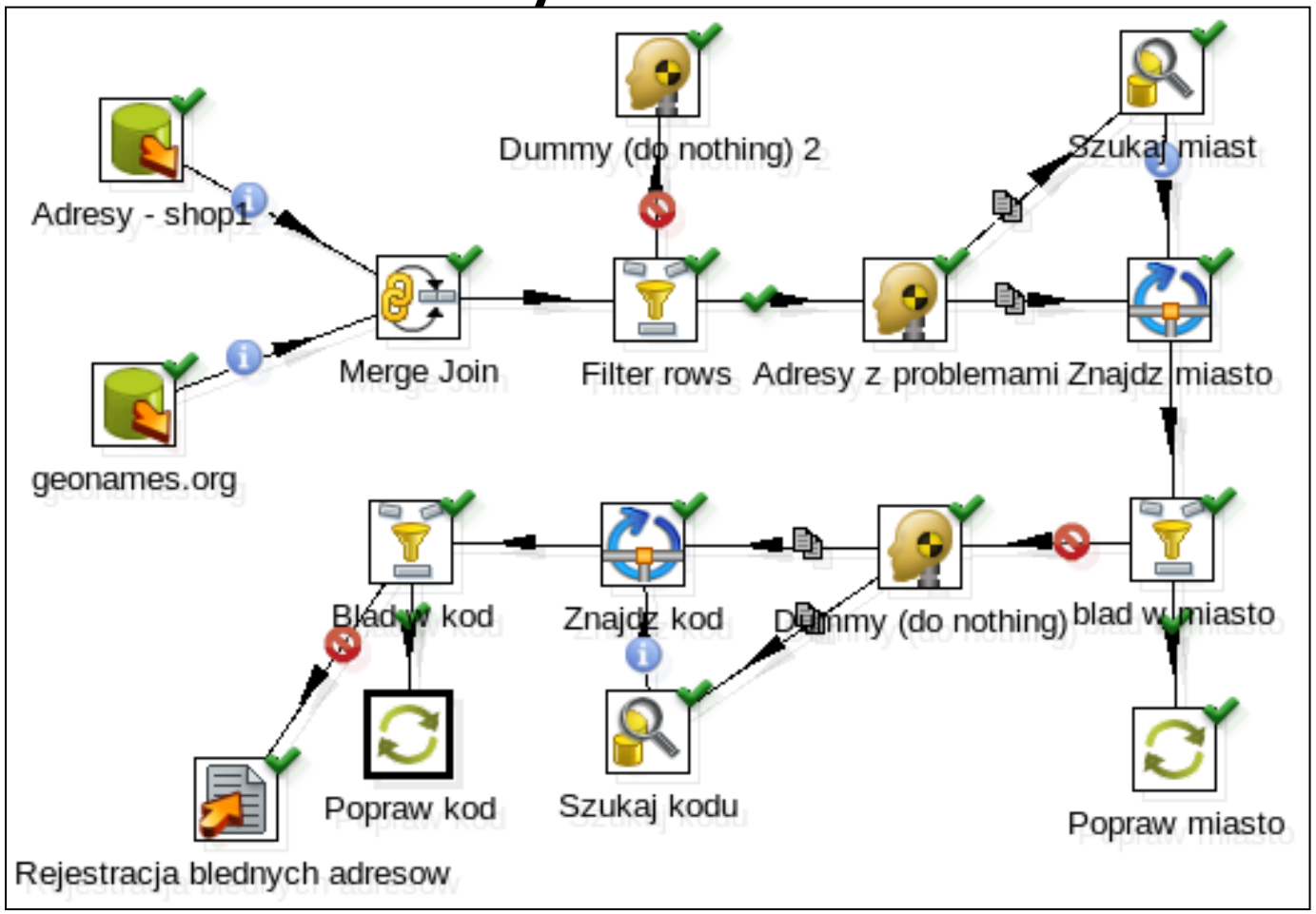
Analiza problemu – profilowanie danych



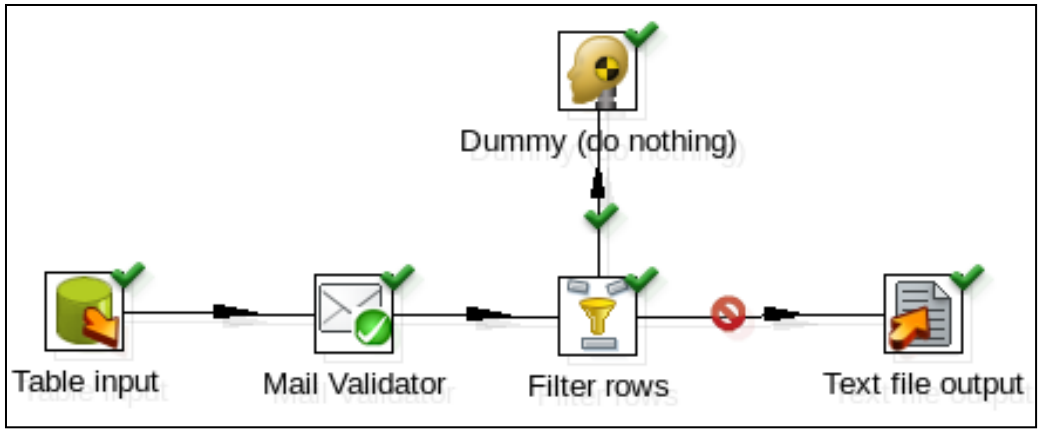
Rows of step: Dummy (do nothing) (11 rows)

#	adres_id	miasto	kod_pocztowy	postal_code	place_name
1	155	Creig	13345		
2	531	Royersford	19466		
3	348	Sharpnes	25183		
4	504	Fortmyers	33967		
5	416	Pataka	47666		
6	572	Oaks	58474		
7	119	Hordville	68866		
8	426	Winston	71856		

Automatyczna naprawa danych adresowych w źródłach

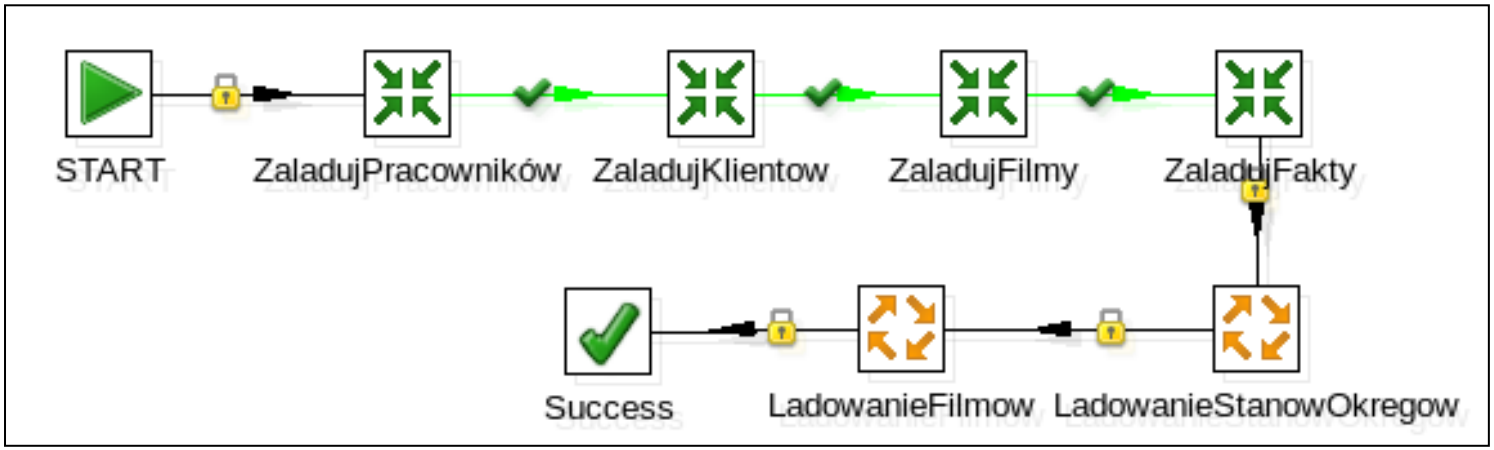


Transformacja weryfikująca poprawność adresów e-mail

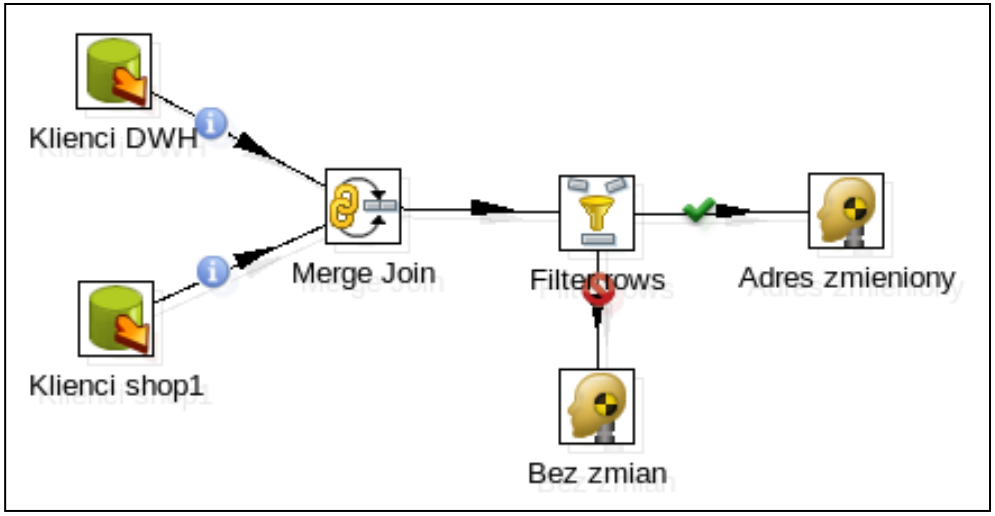



```
klient_id;email;result;Error message
1;MARY.SMITH@sakilacustomer;false;Malformed address MARY.SMITH@sakilacustomer!
2;PATRICIA.JOHNSONsakilacustomer.org;false;Malformed address PATRICIA.JOHNSONsakilacustomer.org!
```

Odświeżenie hurtowni danych – poprawa jakości danych w hurtowni



Analiza zmian wartości atrybutów w wymiarach



Rows of step: Adres zmieniony (3 rows)

#	KL_KLIENT_ID	KL_KOD_POCZTOWY	KL_MIASTO	KL_ULICA	klient_id	miasto	adres	kod_pocztowy
1	285	26209	Snowshoe	1336 Benin City Drive	285	Los Angeles	98 Stara Zagora Boulevard	90001
2	342	65583	Waynesville	1293 Nam Dinh Way	342	Eckert	1192 Tongliao Street	81418
3	399	77850	Concord	734 Bchar Place	399	Vera	953 Hodeida Street	74082

Modyfikacja schematu hurtowni danych na potrzeby CDC type 2

```
SQL> alter table KLIENCI add KL_KLIENT_SID NUMBER(10);

Table altered.
```

```
SQL> create sequence SEQ_KLIENT_SID;

Sequence created.
```

```
SQL> alter table KLIENCI add KL_START_VER TIMESTAMP(6);

Table altered.

SQL> alter table KLIENCI add KL_STOP_VER TIMESTAMP(6);
```

```
SQL> alter table KLIENCI add KL_KLIENT_VER NUMBER(5) DEFAULT 1;

Table altered.
```

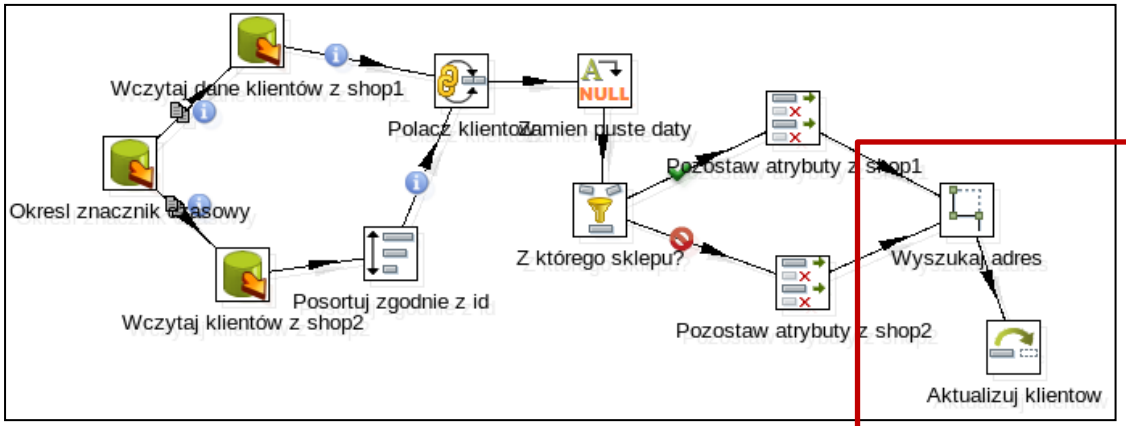
```
SQL> alter table KLIENCI drop primary key cascade;

Table altered.
```

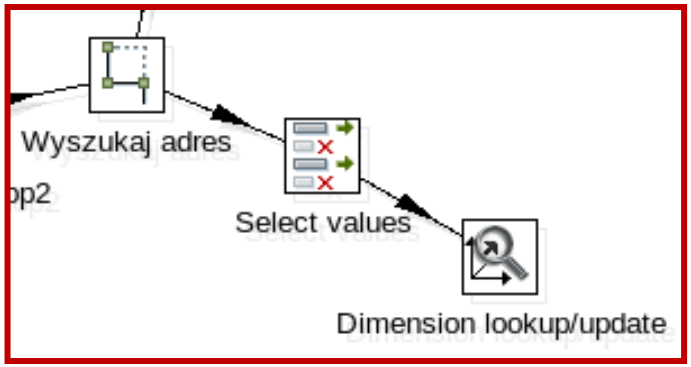
```
SQL> alter table KLIENCI add primary key(KL_KLIENT_SID);

Table altered.
```

```
SQL> update WYPOZYCZENIA
  2  set WY_KLIENT_ID = (select KL_KLIENT_SID from KLIENCI
  3                      where  KL_KLIENT_ID = WY_KLIENT_ID)
  4  ;
```

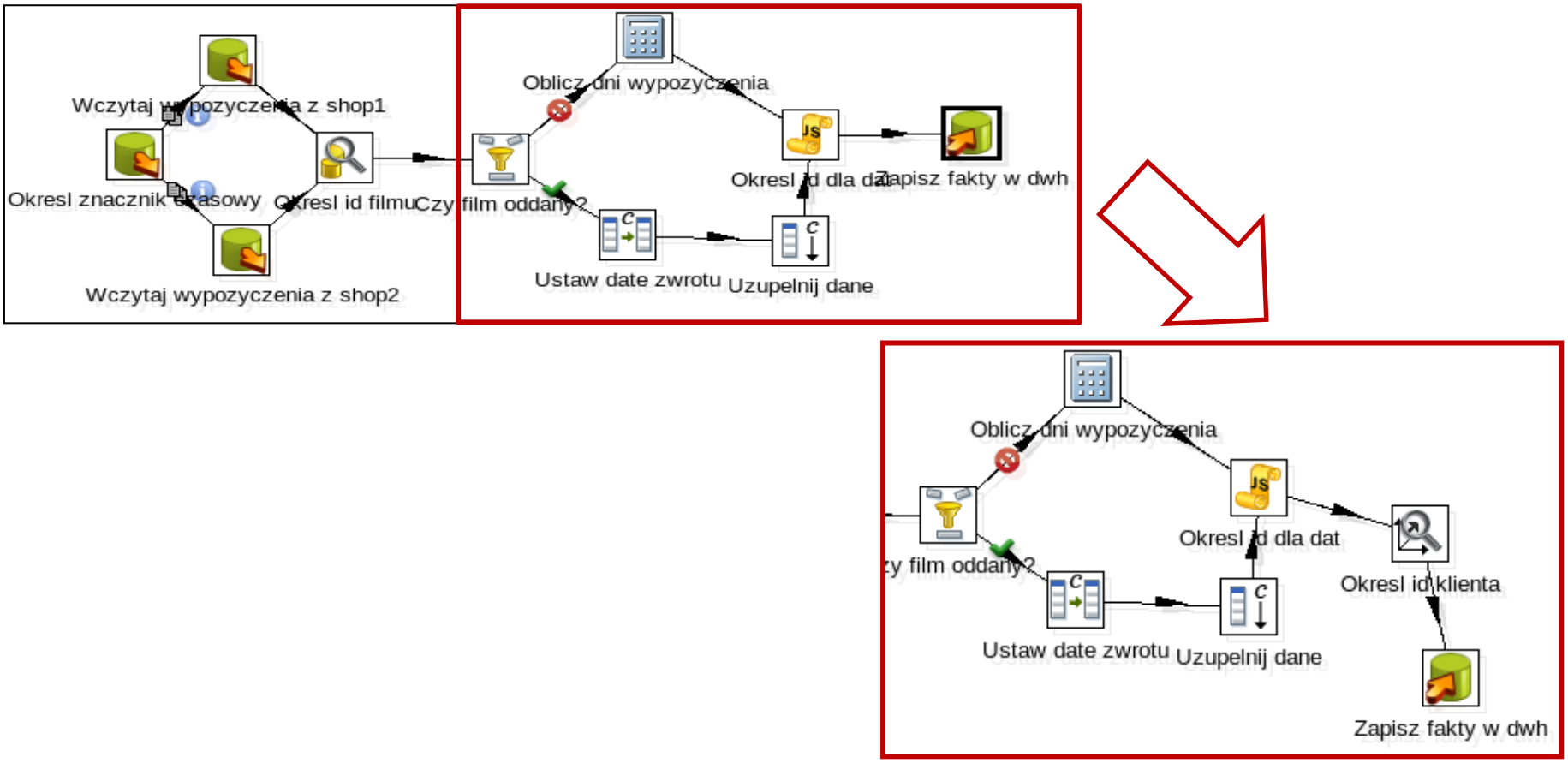
```
SQL> alter table WYPOZYCZENIA rename column WY_KLIENT_ID to WY_KLIENT_SID;

Table altered.

SQL> alter table WYPOZYCZENIA
add foreign key(WY_KLIENT_SID) references KLIENCI(KL_KLIENT_SID);
  2
Table altered.
```

Modyfikacja transformacji wymiaru klient implementująca CDC type 2





Modyfikacja transformacji faktów uwzględniająca CDC type 2



Część 8

TEMATYCZNE HURTOWNIE DANYCH
MASOWE ŁADOWANIE DANYCH

Tematyczne hurtownie danych (ang. *data marts*)

- Tematyczne hurtownie danych tworzone są dla grup osób odpowiedzialnych za konkretny obszar biznesowy.
- W odróżnieniu od hurtowni danych, która ma charakter korporacyjny, tematyczna hurtownia danych wykorzystywana jest zatem przez określony dział, jednostkę.
- Zakres danych tematycznej hurtowni danych również jest ograniczony do zagadnień, będących w sferze zainteresowań wydziału, jednostki.

Tematyczne hurtownie danych cechy i cele

- Łatwy dostęp do danych często wymaganych
- Poprawa czasów odpowiedzi
- Stosunkowo niski koszt implementacji
- Jednoznacznie zdefiniowana grupa końcowych użytkowników – przekłada się to na jednoznaczny cel i zawartość
- Prostsza budowa aplikacji analitycznych – dane ograniczone do tych, które są istotne

Masowe ładowanie danych

- Przez masowe ładowanie danych rozumiemy techniki wydajnego ładowania nowych danych do wnętrza bazy danych
- Dla różnych systemów baz danych masowe ładowanie danych może oznaczać różne odmienne od siebie techniki i mechanizmy

Masowe ładowanie danych
Oracle

- Ścieżka konwencjonalna
  - system analizuje wolne przestrzenie w istniejących już blokach wstawiając nowe wiersze w wolne miejsca znajdujące się wśród istniejących już danych
  - weryfikuje wszystkie ograniczenia integralnościowe
  - zapisuje dane do bloków danych w buforze danych
- Ścieżka bezpośrednia (masowe ładowanie danych)
  - system dołącza wstawiane dane za istniejącymi
  - dane są zapisywane bezpośrednio do plików omijając bufor danych
  - nie jest analizowana wolna przestrzeń w blokach
  - ograniczenia referencyjne są ignorowane
  - istnieje możliwość zrównoleglenia operacji wstawiania wierszy
  - istnieje możliwość wyłączenia zapisów do plików dziennika powtórzeń

Masowe ładowanie danych

MySQL

- Podczas wstawiania wiersza do bazy MySQL należy uwzględnić konieczność: (1) połączenia się, (2) przesłania polecenia do bazy danych, (3) analizy (parsowania) polecenia, (4) wstawienia wiersza, (5) aktualizacji indeksów, (6) zamknięcia
- Ładowanie danych w MySQL można przyspieszyć stosując różne techniki:
  - pojedyncze polecenie z wieloma składowymi VALUES
  - masowe ładowanie danych do tabel MyISAM
    - pojedyncze polecenie
    - ładuje dane z pliku
    - nie aktualizuje indeksów
    - dodatkowa optymalizacja w przypadku ładowania do pustej tabeli
    - możliwość zrównoleglenia operacji wstawiania

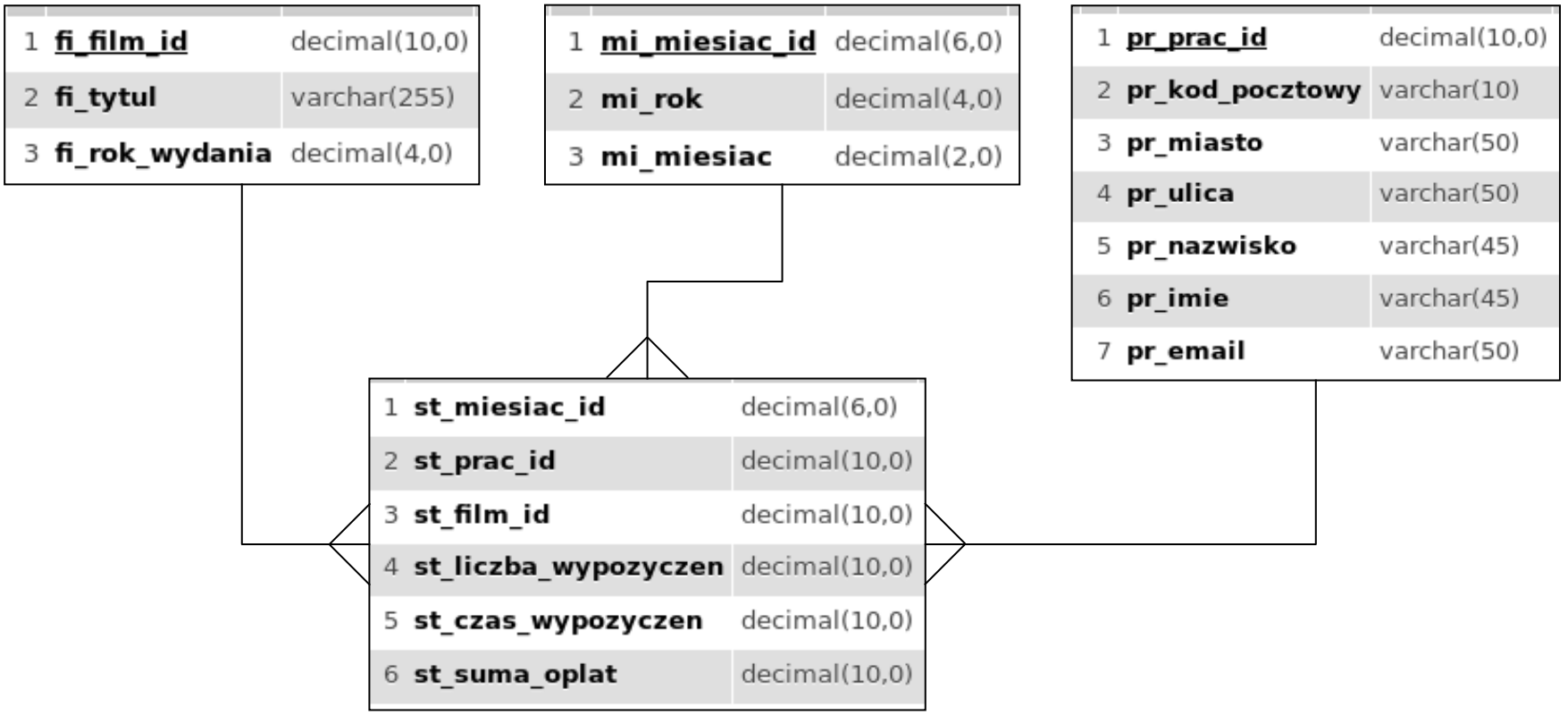
Masowe ładowanie danych
PostgreSQL

- Do masowego ładowania danych w przypadku PostgreSQL służy polecenie COPY
  - pozwala na załadowanie danych bezpośrednio z pliku.
  - nie uruchamia żadnych wyzwalaczy
  - dane z pliku zaczytywane są bezpośrednio przez proces serwera a nie klienta (plik musi znajdować się na maszynie serwera)
- Polecenie copy dostępne w programie psql (wykorzystywane przez *Pentaho DI*) wywołuje polecenie COPY.
  Różnica jest taka, że:
  - odczyt z/do pliku wykonywany jest w tym przypadku przez psql (klienta)
  - nie są wymagane uprawnienia administratora

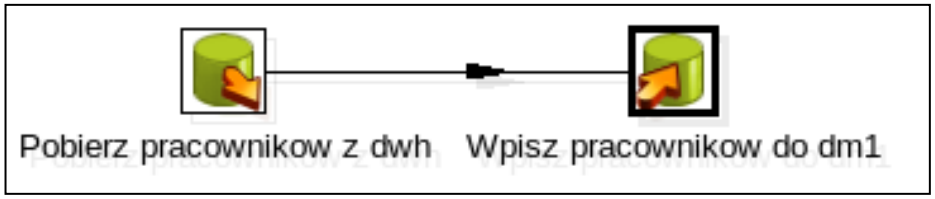
Omówienie ćwiczeń

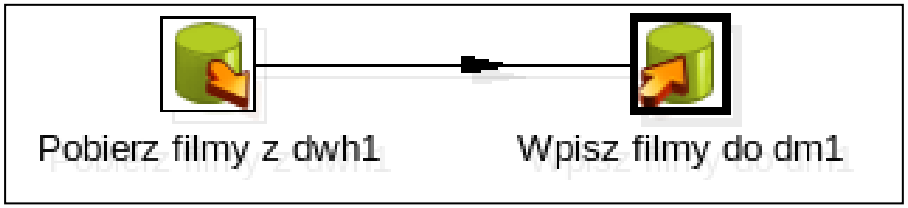
- Tematyczne hurtownie danych
  - Analiza schematu tematycznej hurtowni danych
  - Transformacja wymiarów filmy i pracownicy
  - Transformacja wymiaru dat oraz faktów
  - Rozszerzenie transformacji na drugą tematyczną hurtownię danych
- Zakończenie ćwiczeń
  - Wykorzystanie przykładowej aplikacji analitycznej
  - Uruchomienie zadania odświeżającego hurtownię danych z pomocą linii poleceń
  - Obserwacja zmienionych danych

Analiza schematu tematycznej hurtowni danych

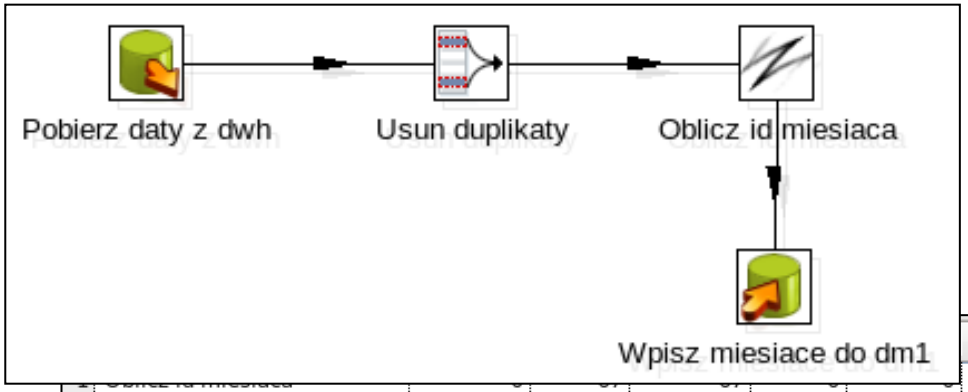


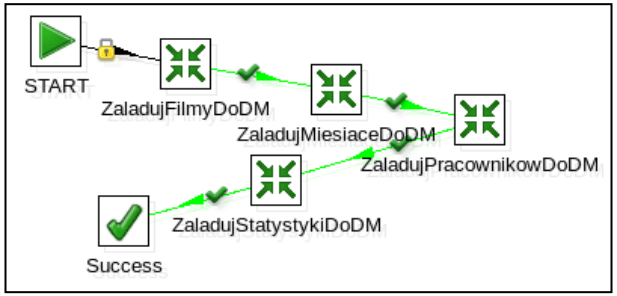
Transformacja wymiarów filmy i pracownicy





Transformacja wymiaru dat oraz faktów

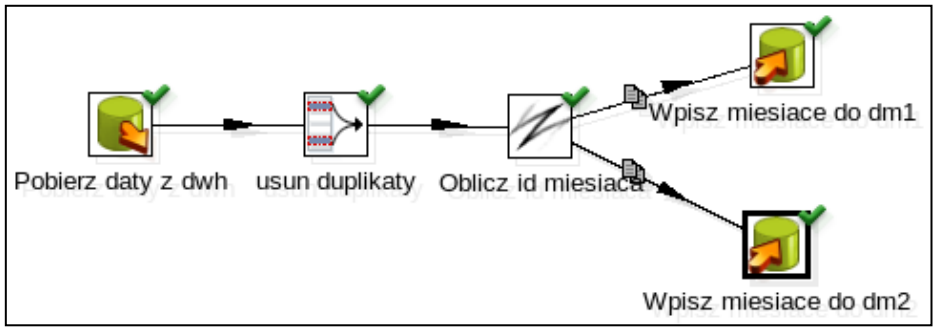


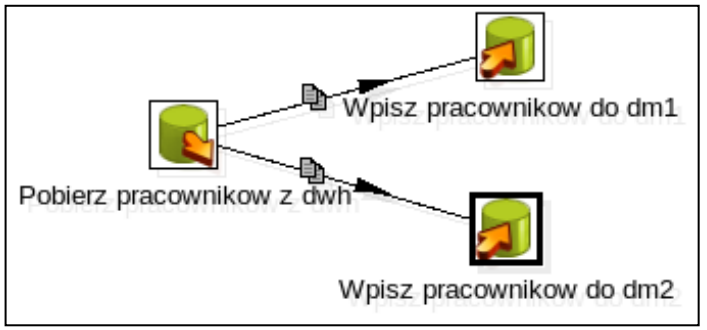


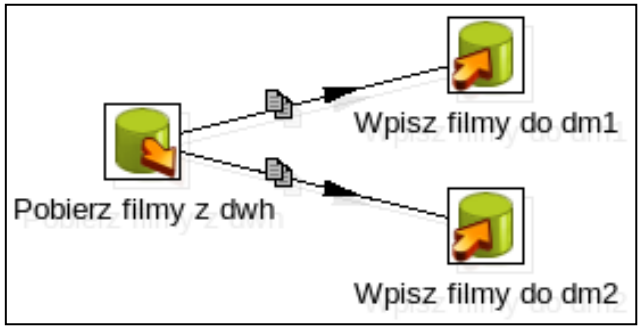
							Updated	Rejected	Errors	Active	Time	Speed (r/s)
							0	0	0	Finished	0.1s	486
2	Wpisz miesiace do dm1	0	67	67	0	67	0	0	0	Finished	0.2s	358
3	usun duplikaty	0	2001	67	0	0	0	0	0	Finished	0.1s	15 045
4	Pobierz daty z dwh	0	0	2001	2001	0	0	0	0	Finished	0.2s	12 428

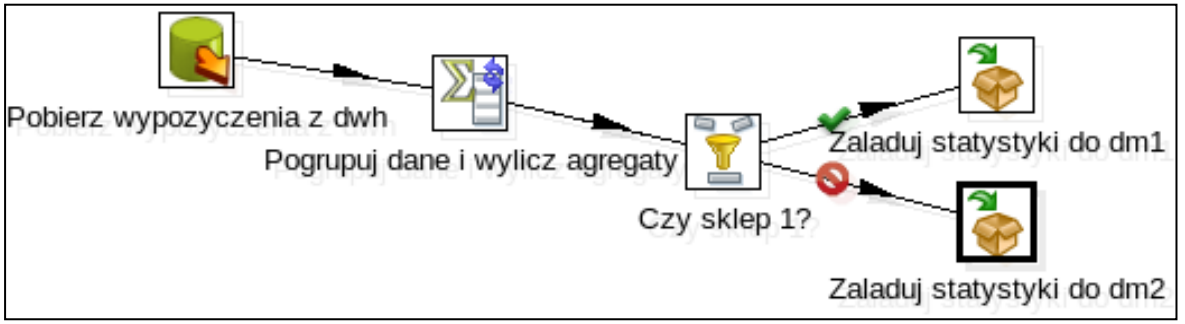
							utput	Updated	Rejected	Errors	Active	Time	Speed (r/s)
							0	0	0	0	Finished	1.8s	8 644
2	Zaladuj statystyki do dm1	0	8186	8186	0	8186	0	0	0	Finished	2.5s	3 316
3	Pobierz wypozyczenia z dwh	0	0	15862	15862	0	0	0	0	Finished	1.3s	12 081

Rozszerzenie transformacji na drugą tematyczną hurtownię danych

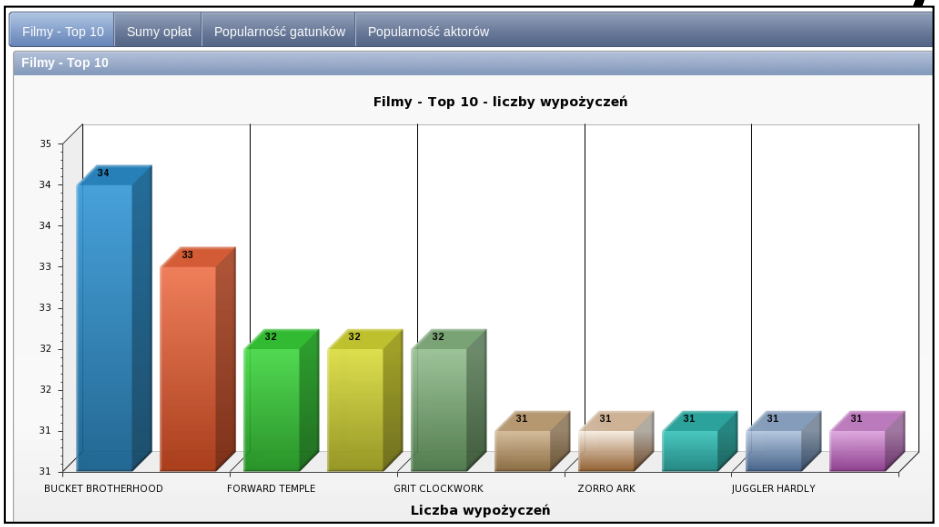


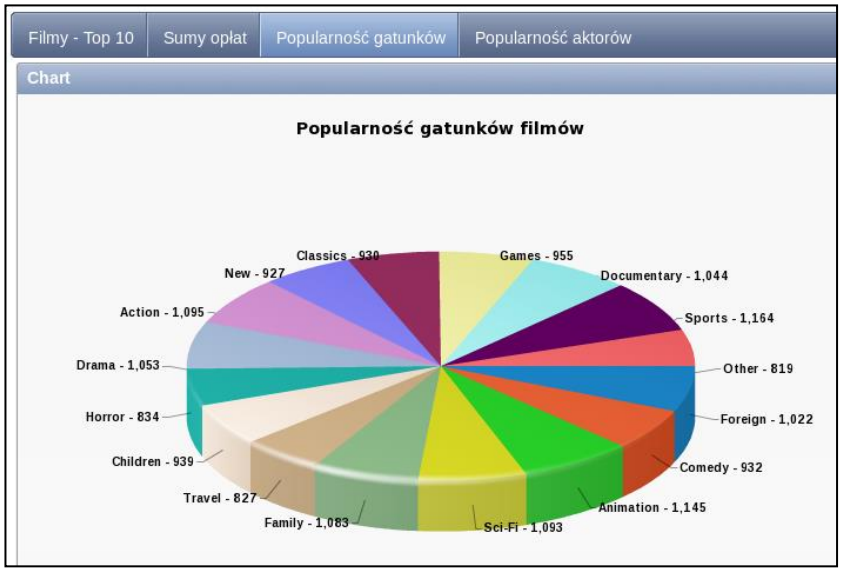






Wykorzystanie przykładowej aplikacji analitycznej





Uruchomienie zadania z linii poleceń i obserwacja zmienionych danych

```
[etl@localhost data-integration]$ ./kitchen.sh -rep:1 -dir:"/wypozyczalnie/shop->dwh" -user:"admin" -pass:"admin" -job=OdswiezDWH
```