

Politechnika Poznańska
Instytut Informatyki

EFEKTYWNOŚĆ ETL i ELT

Na przykładzie bazy odlotów SCGN
oraz danych pogodowych NOAA z lat 1987-2008

Krzysztof Prałat
Tomasz Skrzypczak
Grzegorz Stolarek

1. Spis treści

1.	Spis treści.....	1
2.	Wstęp	3
2.1.	Definicja ETL i ELT	3
2.2.	Cel i zakres pracy	5
2.3.	Struktura pracy	5
3.	Architektura sprzętowa i programowa	7
3.1.	Serwer główny – Source	7
3.2.	Virtual Box 1 - ETL REP.....	8
3.3.	Virtual Box 2 - HD+ELT REP.....	8
4.	Zbiór danych wejściowych.....	9
4.1.	Dane źródłowe.	9
4.2.	Import danych.	9
4.3.	Charakterystyka danych	10
5.	Hurtownia danych	15
6.	Benchmark.....	20
6.1.	Modele ładowania danych	21
6.2.	Workflow WF_ETL	22
6.1.	Workflow WF_ELT	33
7.	Testy	43
7.1.	Procedura testowa	43
7.2.	Etapy testowe.....	43
7.3.	Wyniki.....	44
8.	Wnioski i podsumowanie	47
9.	Literatura i źródła danych.....	47
10.	Spis rysunków	48
11.	Spis tabel	49
12.	Załączniki	50
12.1.	Skrypty SQL.....	50

2. Wstęp

Hurtownia danych jest to:

- uporządkowany tematycznie,
- zintegrowany,
- zawierający wymiar czasowy,
- nieulotny

zbiór danych wspomagających podejmowanie decyzji dla celów strategicznych i analitycznych.

Termin *Hurtownia danych* pojawił się po raz pierwszy w 1970 roku za sprawą Williama Harveya Inmona, amerykańskiego informatyka, jako odpowiedź na wzrastające zapotrzebowanie na wydajne narzędzia służące do analizy danych oraz predykcji. Od tego momentu zagadnienie to przeżywa intensywny rozkwit, powstają nie tylko rozwiązania programowe, ale również autonomiczne architektury sprzętowe, które dzisiaj wspierają problematykę analizy i predykcji trendów w większości korporacji na świecie.

Dużym problemem nie jest tylko przetwarzanie danych, ale przede wszystkim efektywne ich ładowanie oraz integrowanie w stałych odstępach czasowych. Powstały wyspecjalizowane narzędzia takich producentów jak: Oracle, IBM, SAP, Informatica, które obejmują swoim zakresem pełny przepływ od źródeł danych do hurtowni danych.

2.1. Definicja ETL i ELT

Wyróżnia się dwa podstawowe podejścia do ładowania danych do hurtowni, tj. ETL oraz ELT. W praktyce spotkać można również ich wariacje, wykorzystujące najmocniejsze strony każdej z nich.

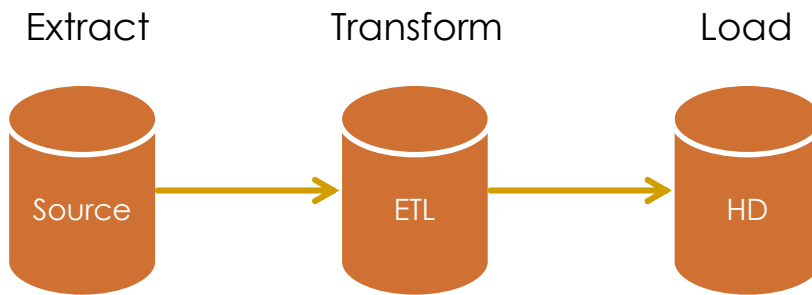
2.1.1. ETL (Extract, Transform, Load)

Proces ETL oznacza pobranie danych ze źródeł, odpowiednie ich przekształcenie i załadowanie do hurtowni danych lub innych baz będących punktem odniesienia dla warstwy aplikacyjnej systemów raportujących.

W praktyce można się również spotkać z określeniem *proces integracji danych*, natomiast warstwa narzędzi ETL znana jest jako platforma integracyjna.

Inne terminy związane z ekstrakcją, transformacją i ładowaniem danych to: migracje danych, zarządzanie danymi, czyszczenie danych, testy jakości danych, synchronizacja danych i konsolidacja danych.

W większości przypadków nadrzędnym celem i korzyścią płynącą z posiadania narzędzia ETL w organizacji jest zarządzanie przepływem danych ze źródłowych systemów OLTP do hurtowni danych i zasilenie tematycznych hurtowni danych (tzw. *data marts*).



Rysunek 2.1 Model architektury ETL

Zalety:

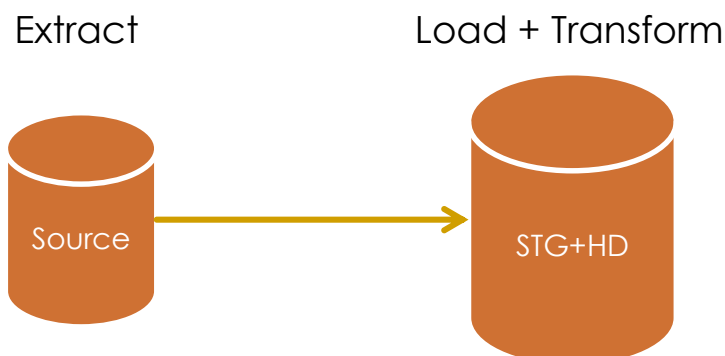
- Nie przetwarzamy i nie przechowujemy w hurtowni zbędnych danych, takich, których nie wykorzystamy później w raportach.
- Więcej możliwości zarządzania przepływem danych.
- Może zrównoważyć i współdzielić obciążenie z SZBD (Systemem Zarządzania Bazą Danych).

Wady:

- Większy nakład pracy przy dokonywaniu zmian w hurtowni, np. dodaniu nowych tabel czy atrybutów (potrzeba tworzenia nowych *workflowów* ETL już od etapu bazy źródłowej).

2.1.2. ELT (Extract, Load, Transform)

Dane są ekstrahowane ze źródła w niezmienionej (lub prawie niezmienionej) postaci do warstwy tzw. *staging area*, gdzie mogą być sprawdzone pod względem integralności. Dane są później ładowane do hurtowni danych, gdzie zachodzi pełna transformacja danych do postaci finalnej.



Rysunek 2.2 Model architektury ELT

Zalety:

- Pełna i zweryfikowana kopia danych źródłowych w hurtowni danych.
- Oddzielenie fazy Extract od Load, a co za tym idzie minimalizowanie potrzeby obciążania bazy źródłowej przy zmianach bądź ładowaniach hurtowni.
- Izolacja faz Load oraz Transform (brak dziedziczonych zależności), co prowadzi do uproszczenia procesów wprowadzania zmian do hurtowni.
- Szybkość przetwarzania w rozwiązaniach typu Teradata (duża część przetwarzania odbywa się w pamięci RAM).

Wady:

- Pełna kopia danych zajmuje dużo miejsca na dyskach hurtowni.
- Transformacje wykorzystują zasoby SZBD, co ma wpływ na generowanie raportów końcowych.
- Ograniczona liczba narzędzi dostępnych na rynku.

2.2. Cel i zakres pracy

Przedmiotem pracy jest porównanie dwóch różnych podejść do zagadnienia ładowania danych: ELT oraz ETL. W celu przeprowadzenia odpowiednich doświadczeń zostały zaprojektowane przepływy ETL oraz ELT dla zdefiniowanych i niezmiennych danych źródłowych. Procedura testowa została podzielona na etapy, po 4 dla ETL i ELT. Z każdym następnym ładowaniem załadowane zostało około 2 razy więcej danych niż w etapie wcześniejszym. Czynnikiem badanym jest czas trwania poszczególnych przepływów - im krótszy, tym lepiej.

2.3. Struktura pracy

Opis projektu podzielony jest na 5 rozdziałów ułożonych chronologicznie, zgodnie z kolejnością wykonywanych zadań.

2.3.1. Rozdział 3 – Architektura sprzętowa

Rozdział 3 to opis architektury sprzętowej i programowej wykorzystanej w realizacji projektu. Zawiera szczegółową listę elementów składowych oraz schematy poglądowe.

2.3.2. Rozdział 4 – Zbiór danych wejściowych

W rozdziale 4 znajduje się: opis danych źródłowych, pochodzenie, znaczenie biznesowe, struktura, opis metadanych.

2.3.3. Rozdział 5 – Hurtownia danych

Rozdział 5 zawiera opis struktury hurtowni danych, znaczenie biznesowe, schemat z dokładnym opisem metadanych.

2.3.4. Rozdział 6 – Benchmark

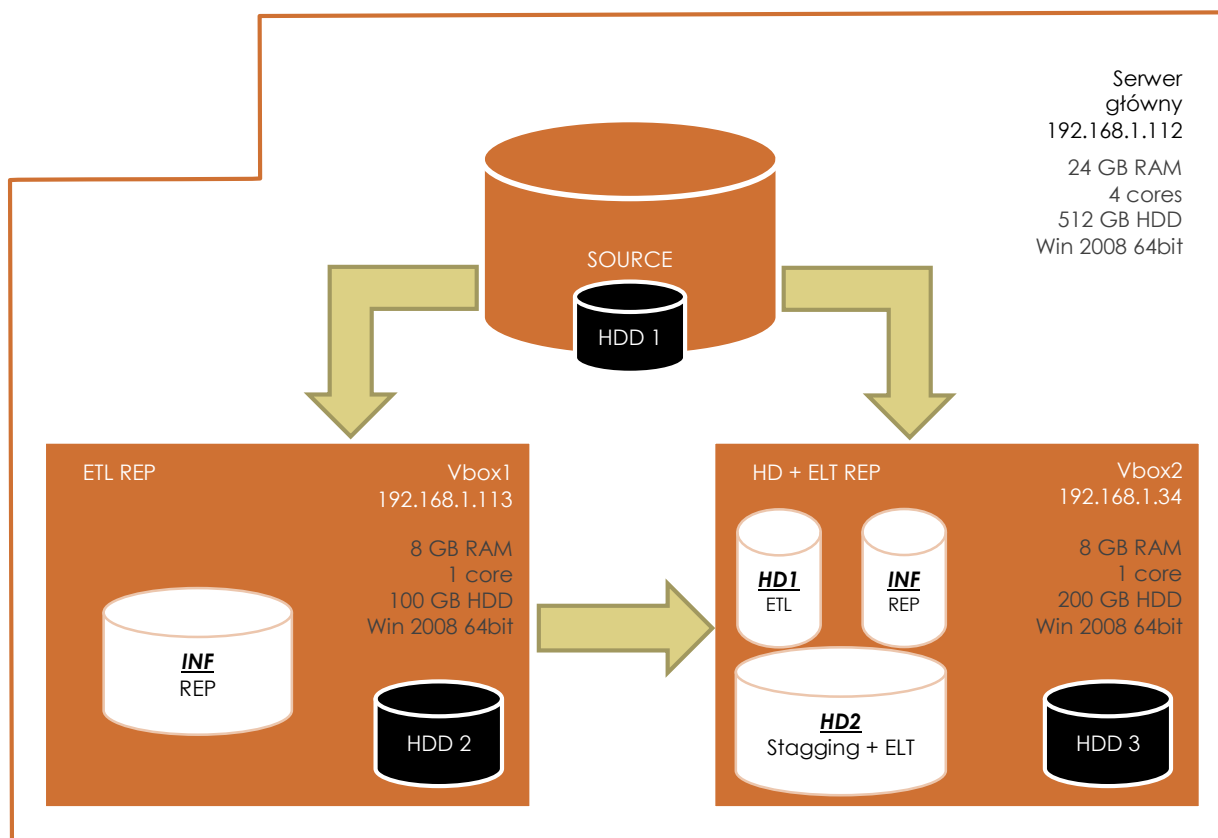
Rozdział 6 przedstawia strukturę przepływu danych źródło – hurtownię, opis poszczególnych *mappingów*, skryptów bazodanowych oraz procedury ładowania.

2.3.5. Rozdział 7 - Testy

W rozdziale 7 opisano przeprowadzone doświadczenia oraz podsumowanie wyników.

3. Architektura sprzętowa i programowa

Wszystkie doświadczenia przeprowadzone zostały na serwerze z 4-rdzeniowym procesorem oraz 24GB pamięci RAM. Został on podzielony logicznie na 3 części: Source, ETL REP oraz HD+ELT REP, umieszczonych na 3 niezależnych dyskach HDD, tak by móc zasymulować rzeczywiste środowisko produkcyjne. Wykorzystana została technologia maszyn wirtualnych VirtualBox firmy Oracle, gdzie umieszczono repozytoria ETL/ELT oraz hurtownie danych.



Rysunek 3.1 Architektura oraz model przepływu

3.1. Serwer główny – Source

Środowisko, na którym znajduje się SZBD Oracle w wersji 11g zawierająca dane źródłowe zostało załadowane wcześniej z plików csv.

ELEMENT	OPIS
Procesor	Intel Core i5-3470 CPU @ 3.20 GHz
Pamięć RAM	24 GB
HDD	WD 500GB
System operacyjny	MS Windows Server 2008 R2 EN Standard 64-bit
Numer IP	192.168.1.112
Nazwa serwera	Leonithehouse

Tabela 3.1 Specyfikacja serwera głównego – Source

3.2. Virtual Box 1 - ETL REP

Na pierwszym serwerze wirtualnym zrealizowanym w oparciu o Oracle VirtualBox zainstalowane zostały SZBD Oracle w wersji 11g oraz oprogramowanie platformy integracyjnej Informatica PowerCenter 9.5.1. Zadaniem tego serwera jest realizacja procesu ETL.

ELEMENT	OPIS
Procesor	2 vcpu
Pamięć RAM	8 GB
HDD	100 GB
System operacyjny	MS Windows Server 2008 R2 PL Standard 64-bit
Numer IP	192.168.1.113
Nazwa serwera	ETL REP

Tabela 3.2 Specyfikacja serwera wirtualnego – ETL REP

3.3. Virtual Box 2 - HD+ELT REP

Na drugim serwerze wirtualnym zrealizowanym w oparciu o Oracle VirtualBox zainstalowane zostały SZBD Oracle w wersji 11g oraz oprogramowanie platformy integracyjnej Informatica PowerCenter 9.5.1. Zadaniem tego serwera jest realizacja procesu ELT oraz utrzymanie hurtowni danych.

ELEMENT	OPIS
Procesor	2 vcpu
Pamięć RAM	8 GB
HDD	200 GB
System operacyjny	MS Windows Server 2008 R2 PL Standard 64-bit
Numer IP	192.168.1.34
Nazwa serwera	HD+ETL REP

Tabela 3.3 Specyfikacja serwera wirtualnego – HD+ETL REP

4. Zbiór danych wejściowych

4.1. Dane źródłowe.

Bazą źródłową dla hurtowni danych jest Oracle 11g. Instancja SOURCE zawiera informacje o lotach samolotów w USA w latach 1987-2008, a dane znajdują się w 6 tabelach:

- SRC_D_AIRPORT,
- SRC_D_PLANE_DATA,
- SRC_D_STATES,
- SRC_D_STATIONS,
- SRC_F_FLIGHTS,
- SRC_F_WEATHER.

Do tabeli SRC_D_AIRPORT zaimportowano informacje o portach lotniczych w USA. Dane o portach lotniczych w USA zostały pobrane ze strony <http://stat-computing.org/dataexpo/2009/airports.csv> w postaci pliku CSV.

Do tabeli SRC_D_PLANE_DATA zaimportowano informacje o samolotach. Dane o samolotach zostały pobrane ze strony <http://statcomputing.org/dataexpo/2009/plane-data.csv> w postaci plików CSV.

Tabela SRC_D_STATES jest to tabela zawierająca kody i nazwy stanów w USA - stworzona na podstawie <ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-states.txt>.

Tabela D_STATIONS jest to tabela zawierająca informacje o stacjach meteorologicznych w USA - stworzona na podstawie <ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ghcnd-stations.txt>

Do tabeli SRC_F_FLIGHTS zaimportowano informacje o lotach samolotów w USA w latach 1987-2008. Dane o lotach samolotów w USA w latach 1987-2008 zostały pobrane ze strony <http://stat-computing.org/dataexpo/2009/the-data.html> w postaci plików CSV.

Do tabeli SRC_F_WEATHER zaimportowano informacje o pogodzie w USA w latach 1987-2008. Zawierają dane badawcze z terenu USA wykonanych w latach 1987-2008 i zostały pobrane ze strony ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/ w postaci plików CSV.

4.2. Import danych.

Do importu danych wykorzystano wbudowaną metodę Import Data i narzędzie SQL Loader, która pozwala na podstawie metadanych utworzyć tabele na wzór plików CSV.

Skrypt dla tabeli SRC_F_WEATHER ma postać:

```
load data
infile 'C:\HD\Source\Weather\1987.csv'
infile 'C:\HD\Source\Weather\1988.csv'
...
infile 'C:\HD\Source\Weather\2007.csv'
```

```
infile 'C:\HD\Source\Weather\2008.csv'
append into table SOURCE.SRC_F_WEATHER
fields terminated by ","
( STATION_ID, DATE, ELEMENT, DATA_VALUE, M_FLAG, Q_FLAG, S_FLAG, OBS_TIME
)
```

gdzie:

- infile – ścieżka do pliku CSV,
- append into table SOURCE. SRC_F_WEATHER – nazwa tabeli do jakiej importujemy dane,
- fields terminated by – separator oddzielający użyty w pliku CSV,
- STATION_ID..OBS_TIME – nazwy kolumn, do których importujemy dane.

4.3. Charakterystyka danych

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
STATION_ID	VARCHAR2(50)	Identyfikator stacji
DATE	VARCHAR2(50)	Identyfikator daty o formacie YYYYMMDD
ELEMENT	VARCHAR2(50)	Identyfikator elementu (typ badania)
DATA_VALUE	VARCHAR2(50)	Wynik badania dla każdego elementu
M_FLAG	VARCHAR2(1)	Measurement Flag
Q_FLAG	VARCHAR2(1)	Flaga Quality
S_FLAG	VARCHAR2(1)	Flaga Source
OBS_TIME	VARCHAR2(50)	Czas obserwacji w formacie godz-min (0700 = 7:00 am)
YEAR	NUMBER	Rok

Tabela 4.1 Struktura tabeli SRC_F_WEATHER

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
YEAR	VARCHAR2(40)	1987-2008
MONTH	VARCHAR2(20)	Numer miesiąca
DAYOFMONTH	VARCHAR2(20)	Numer dnia w miesiącu
DAYOFWEEK	VARCHAR2(10)	Numer dnia tygodnia
DEPTIME	VARCHAR2(40)	Rzeczywisty czas odlotu (local, hhmm)
CRSDEPTIME	VARCHAR2(40)	Planowany czas odlotu (local, hhmm)
ARRTIME	VARCHAR2(40)	Rzeczywisty czas przylotu (local, hhmm)
CRSARRTIME	VARCHAR2(40)	Planowany czas przylotu (local, hhmm)
UNIQUECARRIER	VARCHAR2(20)	Unikalny kod przewoźnika (carrier code)
FLIGHTNUM	VARCHAR2(40)	Numer lotu (flight number)

TAILNUM	VARCHAR2(20)	Numer tail samolotu (plane tail number)
ACTUALELAPSEDTIME	VARCHAR2(20)	Rzeczywisty czas podróży w minutach
CRSELAPSEDTIME	VARCHAR2(20)	Planowany czas podróży w minutach
AIRTIME	VARCHAR2(10)	Czas lotu w minutach
ARRDELAY	VARCHAR2(20)	Opóźnienie przylotu w minutach
DEPDELAY	VARCHAR2(20)	Opóźnienie odlotu w minutach
ORIGIN	VARCHAR2(30)	Kod lotniska IATA źródłowego
DEST	VARCHAR2(30)	Kod lotniska IATA docelowego
DISTANCE	VARCHAR2(30)	Dystans w milach
TAXIIN	VARCHAR2(20)	Taxi in time, in minutes
TAXIOUT	VARCHAR2(20)	Taxi out time, in minutes
CANCELLED	VARCHAR2(10)	Czy lot anulowany
CANCELLATIONCODE	VARCHAR2(20)	Kod (przyczyna) anulowania lotu
DIVERTED	VARCHAR2(10)	Przekierowanie
CARRIERDELAY	VARCHAR2(20)	Opóźnienie przewoźnika w minutach
WEATHERDELAY	VARCHAR2(20)	Opóźnienie z powodu pogody w minutach
NASDELAY	VARCHAR2(20)	Opóźnienie NSA
SECURITYDELAY	VARCHAR2(20)	Opóźnienie z powodu bezpieczeństwa
LATEAIRCRAFTDELAY	VARCHAR2(20)	Ostateczne opóźnienie samolotu
ID	NUMBER	

Tabela 4.2 Struktura tabeli SRC_F_FLIGHTS

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
IATA	VARCHAR2(50)	Kod lotniska IATA
AIRPORT	VARCHAR2(50)	Nazwa portu
CITY	VARCHAR2(50)	Miasto
STATE	VARCHAR2(50)	Stan
COUNTRY	VARCHAR2(50)	Kraj
LAT	VARCHAR2(50)	Szerokość geograficzna
LONGS	VARCHAR2(50)	Długość geograficzna

Tabela 4.3 Struktura tabeli SRC_D_AIRPORT

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
TAILNUM	VARCHAR2(20)	Numer tail samolotu
TYPE	VARCHAR2(20)	Typ
MANUFACTURER	VARCHAR2(30)	Producent
ISSUE_DATE	VARCHAR2(10)	Data produkcji

MODEL	VARCHAR2(20)	Model
STATUS	VARCHAR2(20)	Status
AIRCRAFT_TYPE	VARCHAR2(30)	Typ samolotu
ENGINE_TYPE	VARCHAR2(20)	Typ silnika
YEAR	VARCHAR2(4)	Rok

Tabela 4.4 Struktura tabeli SRC_D_PLANE_DATA

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
CODE	VARCHAR2(2)	Kod stanu
NAME	VARCHAR2(50)	Pełna nazwa stanu

Tabela 4.5 Struktura tabeli SRC_D_STATES

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
ID	VARCHAR2(11)	Identyfikator stacji
LATITUDE	VARCHAR2(10)	Szerokość geograficzna
LONGITUDE	VARCHAR2(10)	Długość geograficzna
ELEVATION	VARCHAR2(2)	Wysokość stacji (w metrach, brak danych = -999.9)
STATE	VARCHAR2(2)	Stan w USA
NAME	VARCHAR2(30)	Nazwa stacji
GSNFLAG	VARCHAR2(3)	Flaga wskazuje, czy dana stacja wchodzi w skład GCOS
HCNFLAG	VARCHAR2(3)	Flaga wskazuje, czy dana stacja wchodzi w skład U.S.HCN
WMOID	VARCHAR2(5)	ID stacji WMO (brak = null)

Tabela 4.6 Struktura tabeli SRC_D_STATIONS

4.3.1. Indeksy

Na potrzeby projektu zostały założone indeksy na największych tabelach z danymi.

TABELA	NAZWA	KOLUMNA	TYP INDEKSU
SRC_D_STATIONS	STAT_ID_UNIQUE_IDX	ID	NORMAL INDEX
	STAT_STATE_BITMAP_IDX	STATE	BITMAP INDEX
SRC_F_FLIGHTS	FL_DEST_BITMAP_IDX	DEST	BITMAP INDEX
	FL_ORIGIN_BITMAP_IDX	ORIGIN	BITMAP INDEX
	FL_TAILNUM_BITMAP_IDX	TAILNUM	BITMAP INDEX
	UNIQUECARRIER_BITMAP_IDX	UNIQUECARRIER	BITMAP INDEX
	FL_YEAR_BITMAP_IDX	YEAR	BITMAP INDEX
	WE_YEAR_BITMAP_IDX	YEAR	BITMAP INDEX
SRC_F_WEATHER	W_DATE_IDX	W_DATE_IDX	NORMAL INDEX
	W_STATION_ID_IDX	STATION_ID	NORMAL INDEX

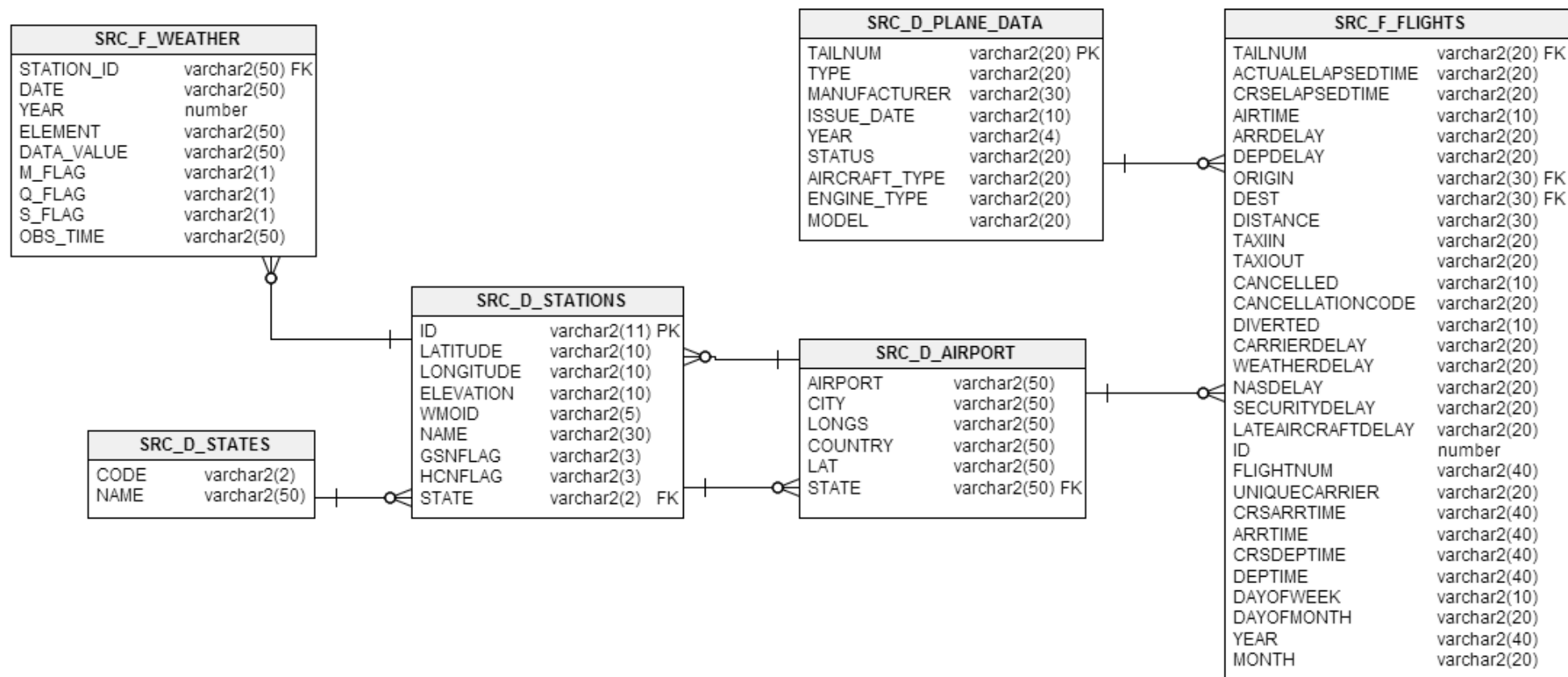
Tabela 4.7 Indeksy

Po imporcie danych w bazie źródłowej mamy 6 tabel: 2 tabele faktów i 4 tabel wymiarów.

Poniżej znajduje się krótka charakterystyka tabel:

TABELA	PLIK	WIELKOŚĆ PLIKU	IŁOŚĆ DANYCH	IŁOŚĆ BLOKÓW
SRC_D_AIRPORT	CSV	244 438 KB	3 376	35
SRC_D_PLANE_DATA	CSV	428 796 KB	5 029	65
SRC_D_STATES	CSV	1.1 KB	73	12
SRC_D_STATIONS	CSV	7.5 MB	91 402	1
SRC_F_FLIGHTS	CSV	12 GB	123 534 969	1 762 612
SRC_F_WEATHER	CSV	21,5 GB	664 962 135	3 862 359

Tabela 4.8 Charakterystyka tabel



Rysunek 4.1 Schemat bazy SOURCE

5. Hurtownia danych

Hurtownia danych składa się z 7 tabel:

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
DWH_ID	NUMBER	Identyfikator
DWH_SOURCE_KEY	VARCHAR2(50)	Unikalny klucz naturalny
DWH_DELETE_FLAG	VARCHAR2(1)	Flaga usunięcia rekordu
DWH_VALID_FROM	DATE	Data wstawienia wiersza
FK_STATION_ID	NUMBER	FK do D_STATIONS
FK_OBS_TIME	NUMBER	FK do kalendarza
STATION_ID	VARCHAR2(50)	Identyfikator stacji
WDATE	VARCHAR2(50)	Identyfikator daty o formacie YYYYMMDD
WELEMENT	VARCHAR2(50)	Identyfikator element (typ badania),
DATA_VALUE	VARCHAR2(50)	Wynik badania dla każdego elementu
M_FLAG	VARCHAR2(1)	Flaga pomiarowa
Q_FLAG	VARCHAR2(1)	Flaga jakości
S_FLAG	VARCHAR2(1)	Flaga źródła
OBS_DATA_TIME	DATE	Czas obserwacji w formacie godz-min (0700 = 7:00)
WYEAR	NUMBER	Rok

Tabela 5.1 Struktura tabeli F_WEATHER

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
DWH_ID	NUMBER	Identyfikator
DWH_SOURCE_KEY	VARCHAR2(50)	Unikalny klucz naturalny
DWH_DELETE_FLAG	VARCHAR2(1)	Flaga usunięcia rekordu
DWH_VALID_FROM	DATE	Data wstawienia wiersza
ID	VARCHAR2(11)	Identyfikator stacji
LATITUDE	VARCHAR2(15)	Szerokość geograficzna
LONGITUDE	VARCHAR2(15)	Długość geograficzna
ELEVATION	VARCHAR2(15)	Wysokość stacji (w metrach, brak danych = -999.9)
STATE_CODE	VARCHAR2(2)	Kod stanu
STATE_NAME	VARCHAR2(50)	Nazwa stanu
NAME	VARCHAR2(30)	Nazwa stacji
GSNFLAG	VARCHAR2(3)	Flaga wskazuje, czy dana stacja wchodzi w skład GCOS
HCNFLAG	VARCHAR2(3)	Flaga wskazuje, czy dana stacja wchodzi w skład U.S.HCN
WMOID	VARCHAR2(11)	ID stacji WMO (brak = null)

Tabela 5.2 Struktura tabeli D_STATIONS

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
DATE_KEY	NUMBER	Wygenerowana na podstawie skryptu załączonego do pracy.
DATE_TIME_START	DATE	
DATE_TIME_END	DATE	
DATE_VALUE	VARCHAR2(11)	
DAY_OF_WEEK_NUMBER	NUMBER	
DAY_OF_WEEK_DESC	VARCHAR2(9)	
DAY_OF_WEEK_SDESC	VARCHAR2(3)	
WEEKEND_FLAG	NUMBER	
WEEK_IN_MONTH_NUMBER	NUMBER	
WEEK_IN_YEAR_NUMBER	NUMBER	
WEEK_START_DATE	DATE	
WEEK_END_DATE	DATE	
ISO_WEEK_NUMBER	NUMBER	
ISO_WEEK_START_DATE	DATE	
ISO_WEEK_END_DATE	DATE	
DAY_OF_MONTH_NUMBER	NUMBER	
MONTH_VALUE	VARCHAR2(2)	
MONTH_DESC	VARCHAR2(9)	
MONTH_SDESC	VARCHAR2(3)	
MONTH_START_DATE	DATE	
MONTH_END_DATE	DATE	
DAYS_IN_MONTH	NUMBER	
LAST_DAY_OF_MONTH_FLAG	NUMBER	
DAY_OF_QUARTER_NUMBER	NUMBER	
QUARTER_VALUE	VARCHAR2(1)	
QUARTER_DESC	VARCHAR2(2)	
QUARTER_START_DATE	DATE	
QUARTER_END_DATE	DATE	
DAYS_IN_QUARTER	NUMBER	
LAST_DAY_OF_QUARTER_FLAG	NUMBER	
DAYS_OF_YEAR_NUMBER	NUMBER	
YEAR_VALUE	VARCHAR2(4)	
YEAR_DESC	VARCHAR2(6)	
YEAR_SDESC	VARCHAR2(4)	
YEAR_START_DATE	DATE	
YEAR_END_DATE	DATE	
DAYS_IN_YEAR	NUMBER	

Tabela 5.3 Struktura tabeli D_CALENDAR

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
DWH_ID	NUMBER	Identyfikator główny
DWH_SOURCE_KEY	VARCHAR2(50)	Unikalny klucz naturalny
DWH_DELETE_FLAG	VARCHAR2(1)	Flaga usunięcia rekordu
DWH_VALID_FROM	DATE	Data wstawienia wiersza
FK_TAILNUM__ID	NUMBER	FK do D_PLANE_DATA
FK_ORIGIN_ID	NUMBER	FK do D_AIRPORT
FK_DEST_ID	NUMBER	FK do D_AIRPORT
FK_DEP_DATE_ID	NUMBER	FK do kalendarza

FK_CRSDEP_DATE_ID	NUMBER	FK do kalendarza
FK_ARR_DATE_ID	NUMBER	FK do kalendarza
FK_CRSARR_DATE_ID	NUMBER	FK do kalendarza
TAILNUM	VARCHAR2(20)	Numer tail samolotu (plane tail number)
ACTUALELAPSEDTIME	VARCHAR2(20)	Rzeczywisty czas podróży w minutach
CRSELAPSEDTIME	VARCHAR2(20)	Planowany czas podróży w minutach
AIRTIME	VARCHAR2(10)	Czas lotu w minutach
ARRDELAY	VARCHAR2(20)	Opóźnienie przylotu w minutach
DEPDELAY	VARCHAR2(20)	Opóźnienie odlotu w minutach
ORIGIN	VARCHAR2(20)	Kod lotniska IATA źródłowego
DEST	VARCHAR2(20)	Kod lotniska IATA docelowego
DISTANCE	VARCHAR2(20)	Dystans w milach
TAXIIN	VARCHAR2(20)	Taxi w czasie, w minutach
TAXIOUT	VARCHAR2(20)	Taxi po czasie, w minutach
CANCELLED	VARCHAR2(20)	Czy lot anulowany
CANCELLATIONCODE	VARCHAR2(20)	Kod anulowania lotu (przyczyna)
DIVERTED	VARCHAR2(20)	Przekierowanie
CARRIERDELAY	VARCHAR2(20)	Opóźnienie przewoźnika w minutach
WEATHERDELAY	VARCHAR2(20)	Opóźnienie z powodu pogody w minutach
NASDELAY	VARCHAR2(20)	Opóźnienie NSA
SECURITYDELAY	VARCHAR2(20)	Opóźnienie ze względów bezpieczeństwa
LATEAIRCRAFTDELAY	VARCHAR2(20)	Ostateczne opóźnienie samolotu
FLIGHTNUM	VARCHAR2(20)	Numer lotu
UNIQUECARRIER	VARCHAR2(20)	Kod linii lotniczej
UNIQUECARRIER_DESC	VARCHAR2(20)	Pełna nazwa linii lotniczej
DEPDATEIME	DATE	Rzeczywisty czas wylotu
CRSDEPDATEIME	DATE	Planowany czas wylotu
ARRDATEIME	DATE	Rzeczywisty czas przylotu
DEPTIME	DATE	Rzeczywisty czas wylotu

Tabela 5.4 Struktura tabeli F_FLIGHTS

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
DWH_ID	NUMBER	Identyfikator
DWH_SOURCE_DAY	VARCHAR2(30)	Unikalny klucz naturalny
DWH_DELETE_FLAG	VARCHAR2(1)	Czy rekord został usunięty
DWH_VALID_FROM	DATE	Data wstawienia wiersza
IATA	VARCHAR2(4)	Kod lotniska IATA
AIRPORT	VARCHAR2(41)	Nazwa portu

CITY	VARCHAR2(33)	Miasto
STATE_CODE	VARCHAR2(2)	Kod stanu
COUNTRY	VARCHAR2(30)	Kraj
STATE_NAME	VARCHAR2(50)	Nazwa stanu
LAT	NUMBER	Szerokość geograficzna
LONGS	NUMBER	Długość geograficzna

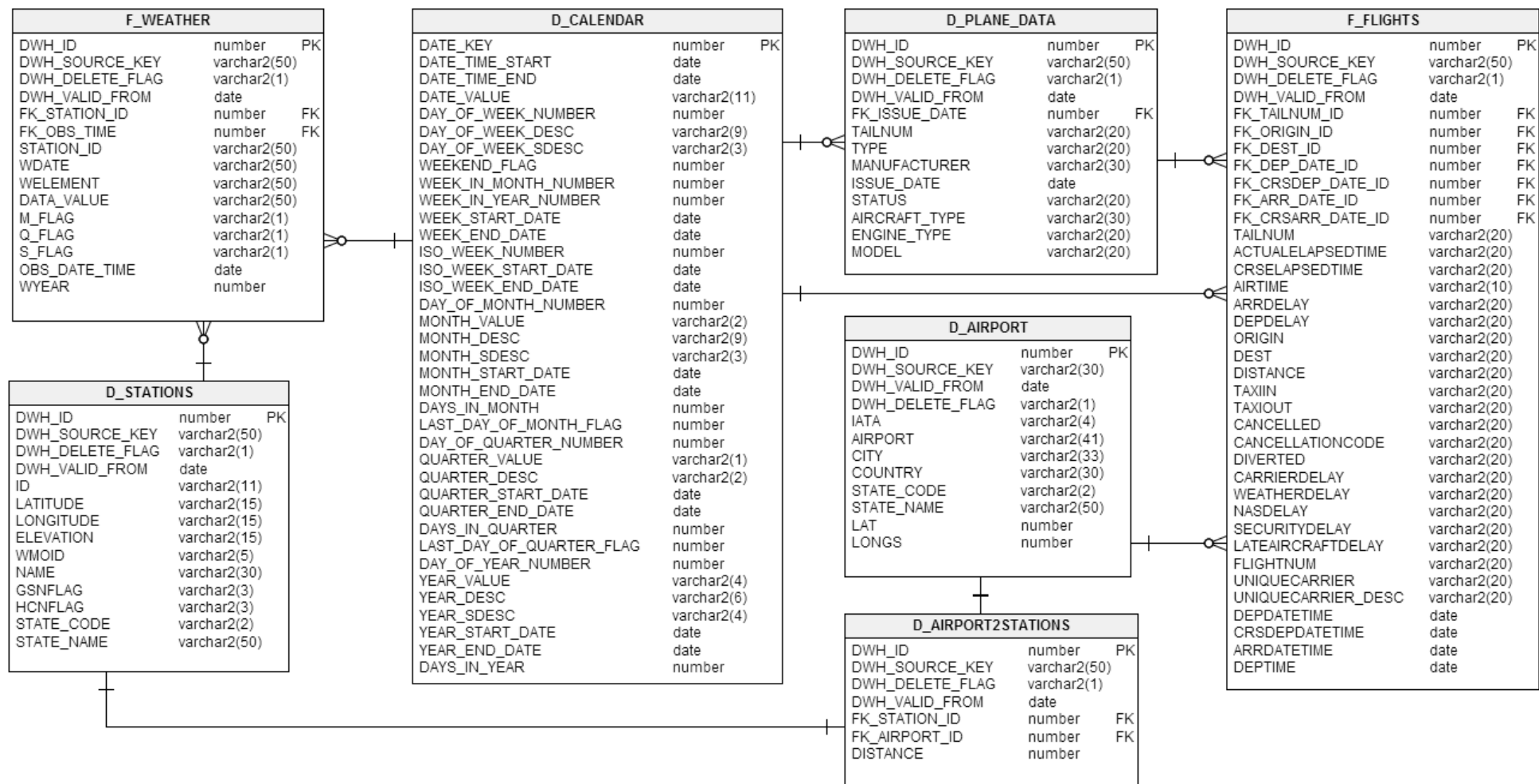
Tabela 5.5 Struktura tabeli D_AIRPORT

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
DWH_ID	NUMBR	Identyfikator
DWH_SOURCE_KEY	VARCHAR2(50)	Unikalny klucz naturalny
DWH_DELETE_FLAG	VARCHAR2(1)	Czy rekord został usunięty
DWH_VALID_FROM	DATE	Data wstawienia wiersza
FK_STATION_ID	NUMBER	Identyfikator stacji
FK_AIRPORT_ID	NUMBER	Identyfikator lotniska
DISTANCE	NUMBER	Dystans w milach

Tabela 5.6 Struktura tabeli D_AIRPORT2STATIONS

ATRYBUT	TYP ATRYBUTU	OPIS ATRYBUTU
DWH_ID	NUMBR	Identyfikator
DWH_SOURCE_KEY	VARCHAR2(50)	Unikalny klucz naturalny
DWH_DELETE_FLAG	VARCHAR2(1)	Czy rekord został usunięty
DWH_VALID_FROM	DATE	Data wstawienia wiersza
FK_ISSUE_DATE	NUMBER	FK do kalendarza
TAILNUM	VARCHAR2(20)	Numer tail samolotu
TYPE	VARCHAR2(20)	Typ
MANUFACTURER	VARCHAR2(30)	Producent
ISSUE_DATE	DATE	Data produkcji
MODEL	VARCHAR2(20)	Model
STATUS	VARCHAR2(20)	Status
AIRCRAFT_TYPE	VARCHAR2(30)	Typ samolotu
ENGINE_TYPE	VARCHAR2(20)	Typ silnika

Tabela 5.7 Struktura tabeli D_PLANE_DATA

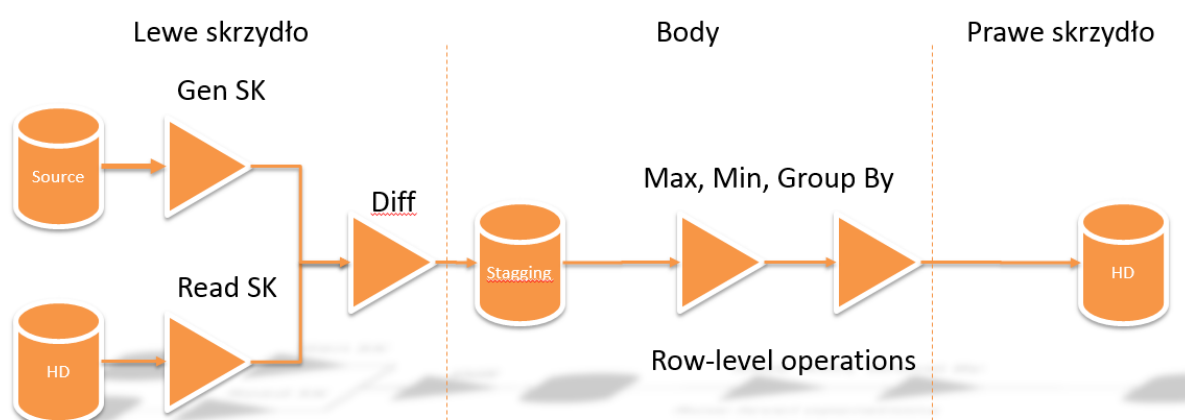


Rysunek 5.1 Schemat hurtowni danych HD

6. Benchmark

Benchmark składa się z dwóch głównych modułów – ETL oraz ELT. Został on stworzony do analizy oraz oceny poszczególnych przebiegów danych, tak by wykazać wyższość jednego podejścia nad drugim. Oba przebiegi korzystają z jednego źródła danych, tj. bazy SOURCE oraz ładują dane do 2 różnych instancji, HD1 oraz HD2 bazy HD. Zależność przebiegów od siebie powoduje, iż testy nie mogą odbywać się jednocześnie, co zostało uwzględnione w procedurze testowej.

Wszystkie przebiegi zostały wykonane na wzór przebiegu Balanced Butterflies, gdzie proces jest podzielony na 3 podstawowe etapy: lewe skrzydło, body oraz prawe skrzydło.



- **Lewe skrzydło**
Etap, w którym dane pobierane są ze źródła, czyszczone, przekształcane i finalnie ładowane do bazy STG, pierwszej nieulotnej instancji w przebiegu.
- **Body**
Nieulotna instancja bazy danych, gromadząca dane przekształcone w lewym skrzydle. Uniezależnia hurtownie danych od źródła (w przypadku awarii, bądź błędu nie ma potrzeby ponownego obciążania źródła)
- **Prawe skrzydło**
Instancja przechowująca dane służące do raportowania i analiz - hurtownia danych. Gromadzi dane uzyskane z STG podczas całego procesu ETL.

W projekcie procesy ładowania są zarządzane poprzez workflowy (dla ETL → WF_ETL oraz dla ELT → WF_EL - podzielone na worklety), które są odpowiedzialne za ładowanie poszczególnych tablic. Dokonują tego w 3 etapach:

- I. Ładowanie ze źródła to tabeli tymczasowej w STG
- II. Ładowanie z tabeli tymczasowej w STG to głównej tabeli STG
- III. Ładowanie z STG do hurtowni danych

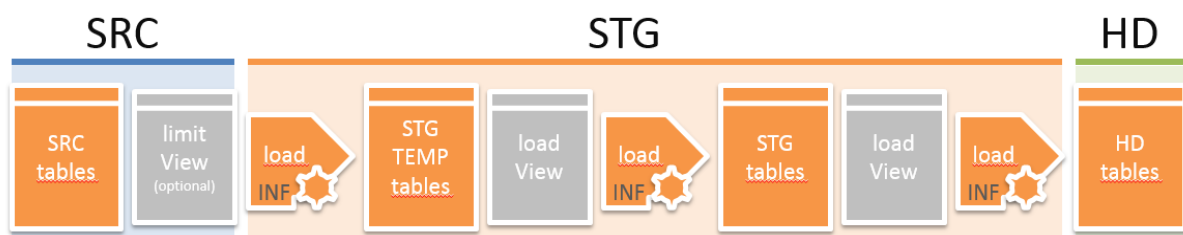
Do ładowania danych została wykorzystana aplikacja PowerCenter Informatica, ponieważ umożliwia ona precyzyjne zarządzanie przepływami oraz mierzenie czasów ładowania i ilości afektowanych wierszy.

6.1. Modele ładowania danych

6.1.1. Model ELT

Agregacja danych odbywa się w silniku RDBMS, a Informatica wykorzystywana jest tylko do ładowania danych punkt-punkt. Realizowane jest to poprzez stworzenie widoków ładujących, transformujących dane do pożądanego kształtu.

1. W pierwszym kroku dane są ładowane ze źródła do tymczasowej tabeli w STG, bez żadnych zmian. Dla tabel F_FLIGHT oraz F_WEATHER zostały utworzone widoki to sterowania ilością danych wejściowych, co ma związek z procedurą testową.
2. Na drugim etapie, w Staggingu, dane są przekształcane do końcowej postaci. Na każdą tablicę źródłową przypadają 2-3 tablice Stagging. Rozbicie na więcej tablic niekiedy okazuje się konieczne ze względu na bardzo ograniczone zasoby sprzętowe.
3. Miejsmem docelowym każdego ładowania jest hurtownia danych (HD), gdzie przechowywane są nieulotnie wyniki wszystkich transformacji w STG. Po każdym ładowaniu wszystkie indeksy w hurtowni są przebudowywane.

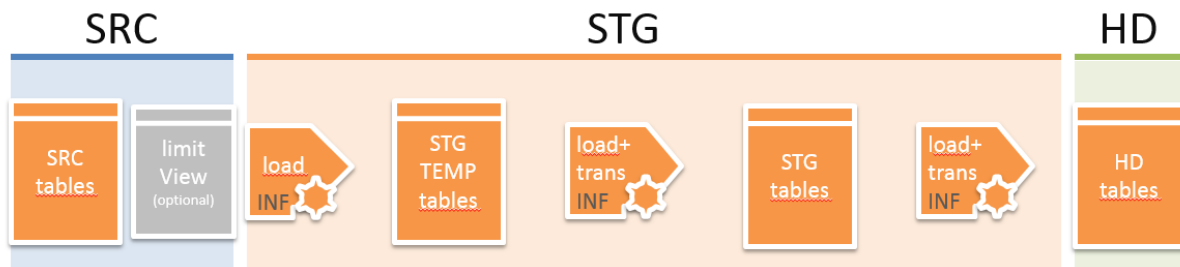


Rysunek 6.1 Model ładowania ELT

6.1.2. Model ETL

W modelu ETL zadanie transformacji danych leży po stronie Informatici. Nie wykorzystuje się przy tym podejściu widoków lecz całą logikę umieszcza się w poszczególnych mapowaniach oraz workletach, składających się na przepływ danych.

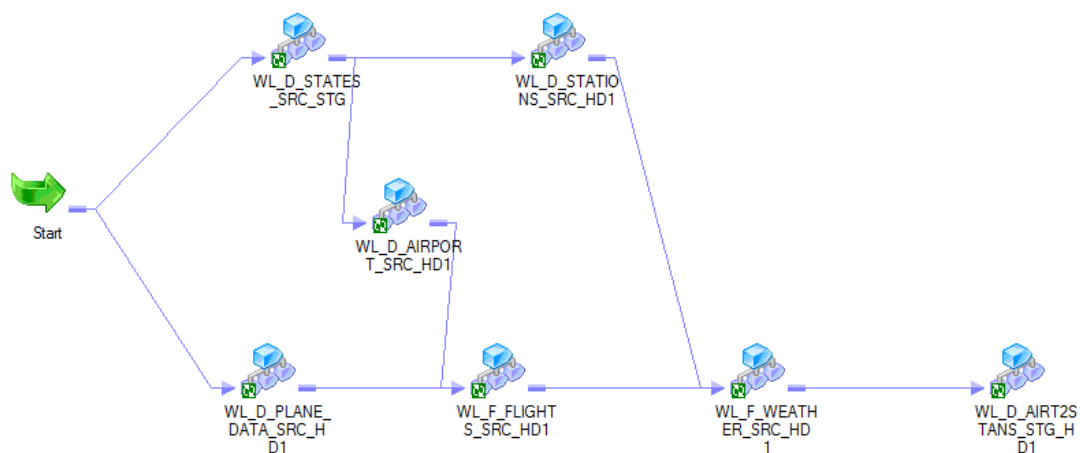
1. Na pierwszym etapie następuje ładowanie danych bez żadnych przekształceń, ze źródła do bazy Staging umieszczonej na tym samym serwerze, co repozytoria Informatici.
2. W bazie Staging następują wszystkie przekształcenia, tj. agregacje, grupowania, czyszczenie danych, wyliczanie delty. Realizowane są one poprzez workflowy, wewnętrzne struktury narzędzia ETL.
3. Po wszystkich przekształceniach dane ładowane są do hurtowni danych (HD).



Rysunek 6.2 Model ładowania ETL

6.2. Workflow WF_ETL

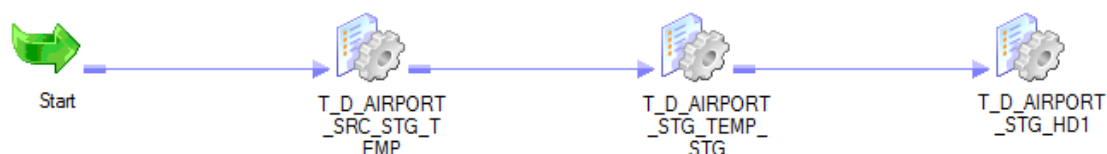
Workflow składa się z 7 workletów odpowiadających ładowaniu 7 tablic ze źródła SOURCE do hurtowni HD1. W następnych podpunktach został sporządzony dokładny opis każdego z nich. Diagram przedstawia zależności pomiędzy poszczególnymi workletami, tj. np. WL_F_FLIGHTS_SRC_HD1 wystartuje dopiero wtedy, kiedy zostaną załadowane 3 inne worklety: WL_D_AIRPORT_SRC_HD1, WL_D_STATES_SRC_STG oraz WL_D_PLANE_DATA_SRC_HD1.



Rysunek 6.3 Workflow WF_ETL

6.2.1. WL_D_AIRPORT_SRC_HD1

Worklet ładujący dane tabeli D_AIRPORT.



Rysunek 6.4 Worklet WL_D_AIRPORT_SRC_HD1

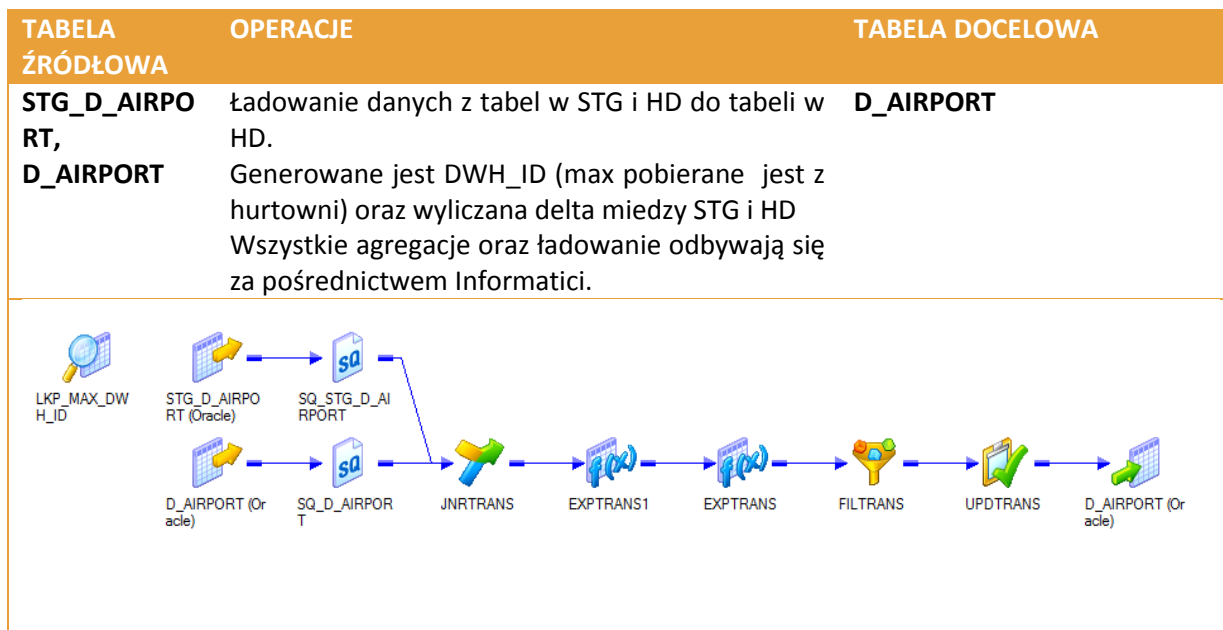
1. D_AIRPORT_SRC_STG_TEMP

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
SRC_D_AIRPORT	Ładowanie danych z tabeli źródłowej do tabeli tymczasowej w STG. Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	STG_D_AIRPORT_TEMP

2. D_AIRPORT_STG_TEMP_STG

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_AIRPORT_TEMP, STG_D_AIRPORT	Ładowanie danych z tabel tymczasowych w STG do tabeli głównej STG. Wyliczany jest SKEY, delta, usuwane duplikaty. Dane są czyszczone oraz przekształcane do postaci końcowej. Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	STG_D_AIRPORT

3. D_AIRPORT_STG_HD1



6.2.2. WL_D_AIRT2STANS_STG_HD1

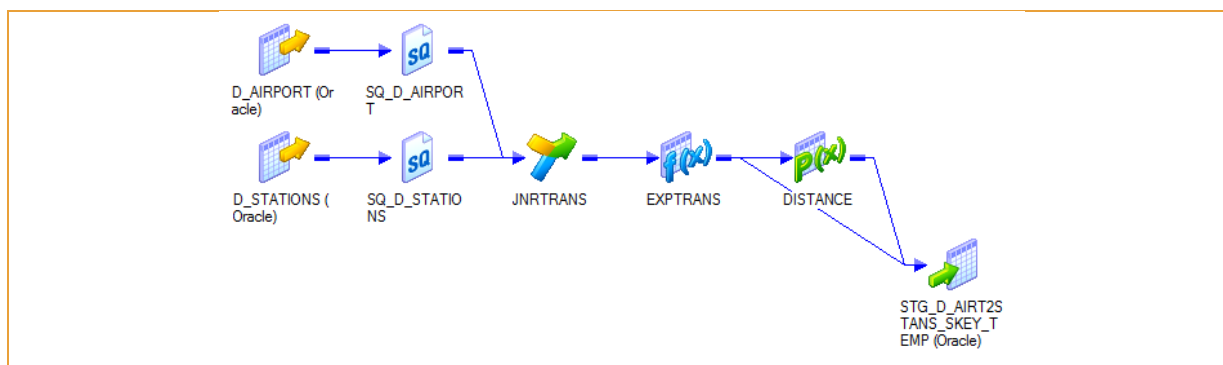
Worklet ładujący dane tabeli D_AIRPORT2STATIONS.



Rysunek 6.5 Worklet WL_D_AIRT2STANS_STG_HD1

1. STG_D_AIRPORT2STATIONS_STG_TEMP_STG_SKEY

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
D_AIRPORT, D_STATIONS	Ładowanie danych z tabel w HD do tabeli w STG. Wyliczany jest SKEY oraz usunięte duplikaty. Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	STG_D_AIRT2STANS_SKEY_TEMP

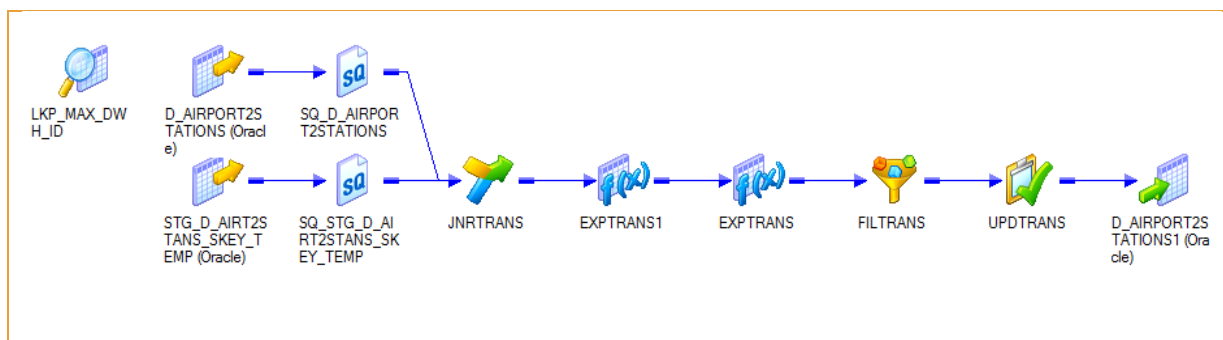


2. STG_D_AIRPORT2STATIONS_STG_SKEY_STG

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_AIRPORT2STATIONS_STG_SKEY	Ładowanie danych z tabel w STG_SKEY do tabeli w STG. Dane są transformowane do postaci końcowej. Wyliczany jest MIN dystans pomiędzy lotniskami i stacjami pogodowymi, do tabeli wstawiane są klucze związane tylko z najkrótszą drogą między punktami. Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	STG_D_AIRPORT2STATIONS_STG_SKEY
<pre> graph LR STG_D_AIRPORT2STATIONS_STG_SKEY[STG_D_AIRPORT2STATIONS_STG_SKEY (Oracle)] --> SQ_STG_D_AIRPORT2STATIONS_STG_SKEY_TEMP[SQ_STG_D_AIRPORT2STATIONS_STG_SKEY_TEMP] SQ_STG_D_AIRPORT2STATIONS_STG_SKEY_TEMP --> maxval[maxval] maxval --> STG_D_AIRPORT2STATIONS_STG_SKEY_TARGET[STG_D_AIRPORT2STATIONS_STG_SKEY (Oracle)] </pre>		

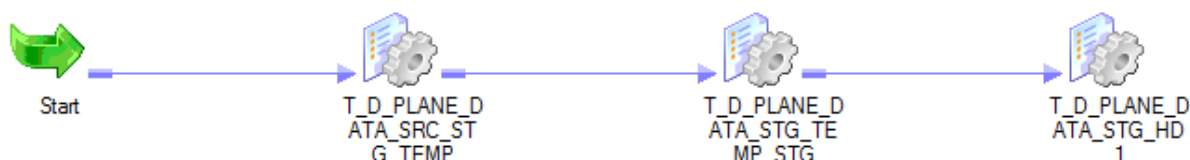
3. STG_D_AIRPORT2STATIONS_STG_HD1

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_AIRPORT2STATIONS_STG_SKEY_TEMP, D_AIRPORT2STATIONS	Ładowanie danych z tabel w STG i HD do tabeli w HD. Dane są transformowane, wyliczana jest delta. Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	D_AIRPORT2STATIONS



6.2.3. WL_D_PLANE_DATA_SRC_HD1

Worklet ładujący dane tabeli D_PLANE_DATA.



Rysunek 6.6 Worklet WL_D_PLANE_DATA_SRC_HD1

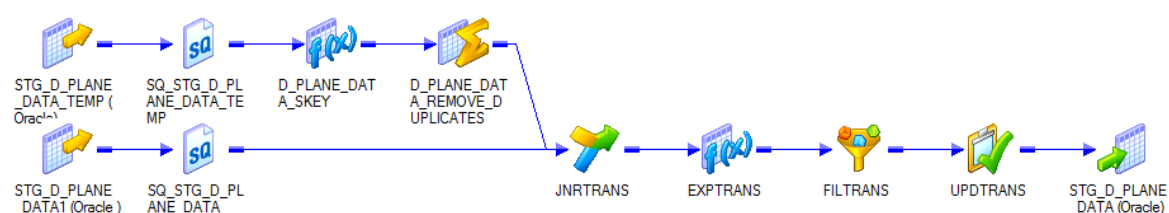
1. D_PLANE_DATA_SRC_STG_TEMP

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
SRC_D_PLANE_DATA	Ładowanie danych z tabeli źródłowej do tabeli tymczasowej w STG. Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatyci.	STG_D_PLANE_DATA_TEMP

2. D_PLANE_DATA_STG_TEMP->STG

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_PLANE_DATA_TEMP, STG_D_PLANE_DATA	Ładowanie danych z tabel tymczasowych w STG do tabeli głównej STG. Wyliczany jest SKEY, delta, usuwane duplikaty. Dane są czyszczone oraz przekształcane do postaci końcowej.	STG_D_PLANE_DATA

Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.



3. D_PLANE_DATA_STG->HD1

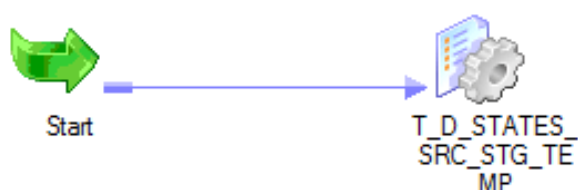
TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_PLANE_DATA, D_PLANE_DATA_A	<p>Ładowanie danych z tabel w STG i HD do tabeli w HD.</p> <p>Generowane jest DWH_ID (max pobierane jest z hurtowni) oraz wyliczana delta między STG i HD</p> <p>Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.</p>	D_PLANE_DATA

```

graph LR
    A[STG_D_PLANE_DATA (Oracle)] --> B[SQ]
    C[D_PLANE_DATA_A (Oracle)] --> D[SQ]
    B --> E[JNRTRANS]
    D --> E
    E --> F[EXPTRANS1]
    F --> G[EXPTRANS]
    G --> H[FILTRANS]
    H --> I[UPDTRANS]
    I --> J[D_PLANE_DATA_A (Oracle)]
  
```

6.2.4. WL_D_STATES_SRC_STG

Worklet ładujący dane tabeli D_STATES, ale tylko do STG. Następnie tablica integrowana jest z STG_D_AIRPORT oraz STG_D_STATIONS.



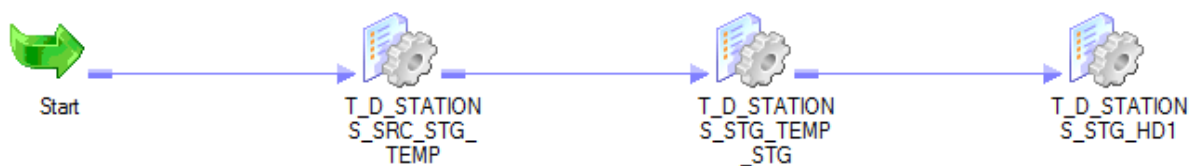
Rysunek 6.7 Worklet WL_D_STATES_SRC_STG

1. D_STATES_SRC_STG_TEMP

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
SRC_D_STATES	Ładowanie danych z tabeli źródłowej do tymczasowej w STG. Dane z tej tabeli są integrowane z tabelami STG_D_AIRPORT oraz STG_D_STATIONS Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	STG_D_STATES_TEMP
<pre> graph LR A[SRC_D_STATES (Oracle)] --> B[SQ_SRC_D_STATES] B --> C[STG_D_STATES_TEMP (Oracle)] </pre>		

6.2.5. WL_D_STATIONS_SRC_HD1

Worklet ładujący dane tabeli D_STATIONS.

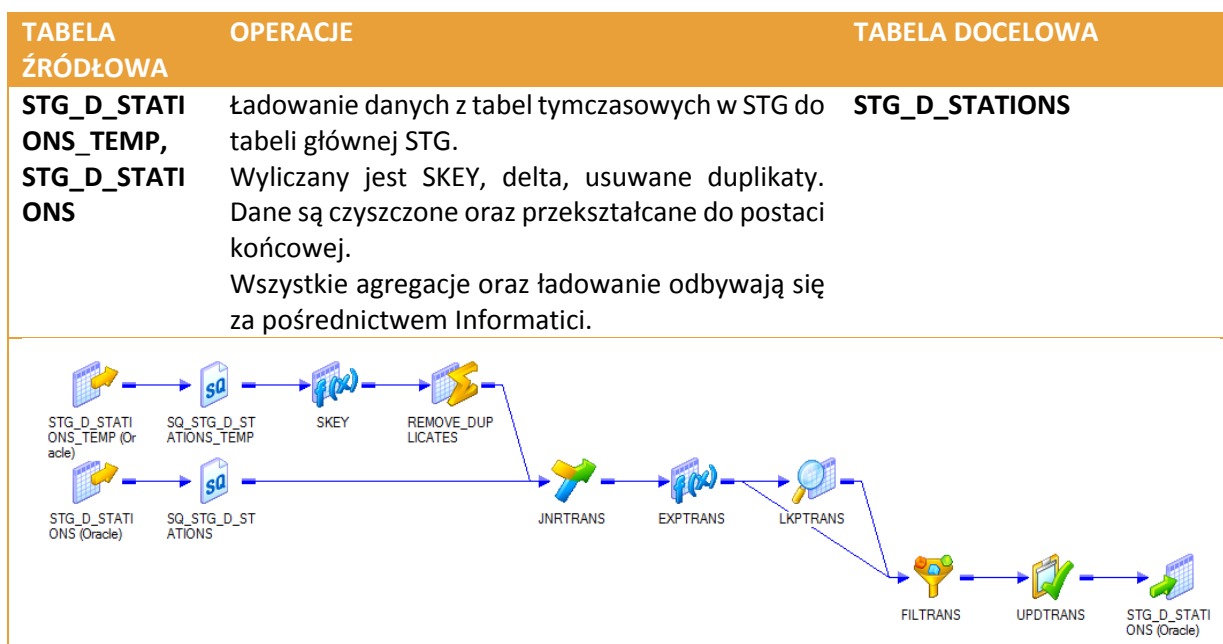


Rysunek 6.8 Worklet WL_D_STATIONS_SRC_HD1

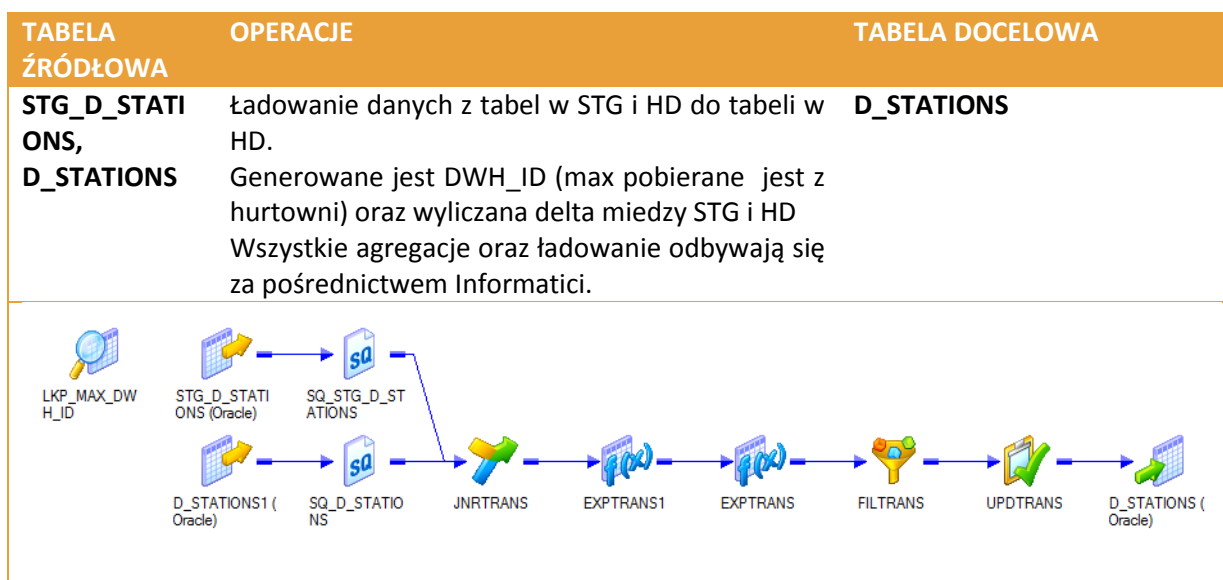
1. D_STATIONS_SRC_STG_TEMP

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
SRC_D_STATIONS	Ładowanie danych z tabeli źródłowej do tabeli tymczasowej w STG. Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	STG_D_STATIONS_TEMP
<pre> graph LR A[SRC_D_STATIONS (Oracle)] --> B[SQ_SRC_D_STATIONS] B --> C[STG_D_STATIONS_TEMP (Oracle)] </pre>		

2. D_STATIONS_STG_TEMP_STG

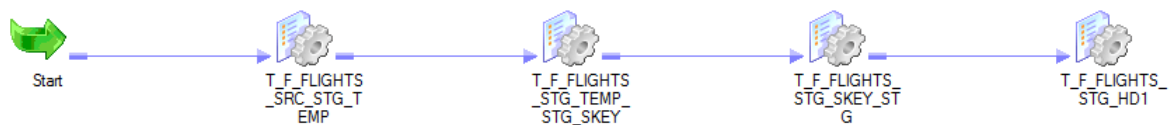


3. D_STATIONS_STG_HD1



6.2.6. WL_F_FLIGHTS_SRC_HD1

Worklet ładujący dane tabeli F_FLIGHTS.



Rysunek 6.9 Worklet WL_F_FLIGHTS_SRC_HD1

1. F_FLIGHT_SRC_STG_TEMP

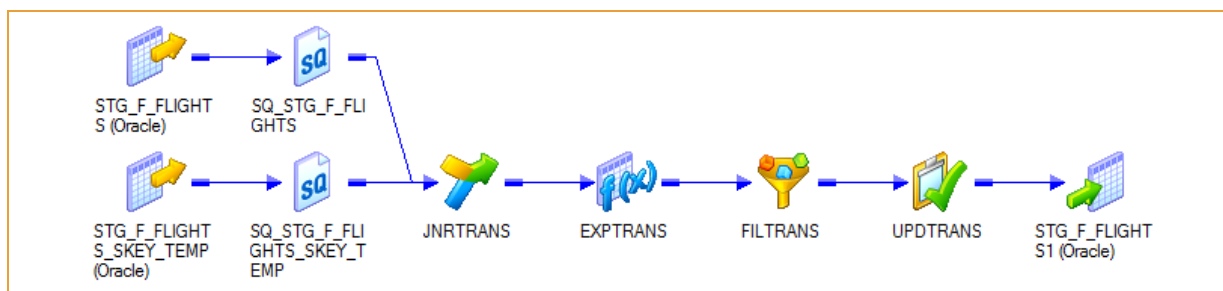
TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
SRC_V_F_FLIGHTS	Ładowanie danych z widoku źródłowego do tabeli tymczasowej w STG. Widok został stworzony do sterowania ilością danych wejściowych. Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	STG_F_FLIGHTS_TEMP

2. F_FLIGHT_STG_TEMP_SKEY

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_F_FLIGHTS_TEMP, STG_F_FLIGHTS	Ładowanie danych z tabel tymczasowych w STG do tabeli głównej STG. Wyliczany jest SKEY, delta, usuwane duplikaty. Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	STG_F_FLIGHTS_SKEY_TEMP

3. F_FLIGHT_STG_SKEY_STG

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_F_FLIGHTS_TEMP, STG_F_FLIGHTS	Ładowanie danych z tabel tymczasowych w STG do tabeli głównej STG. Dane są czyszczone oraz transformowane do postaci końcowej. Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	STG_F_FLIGHTS

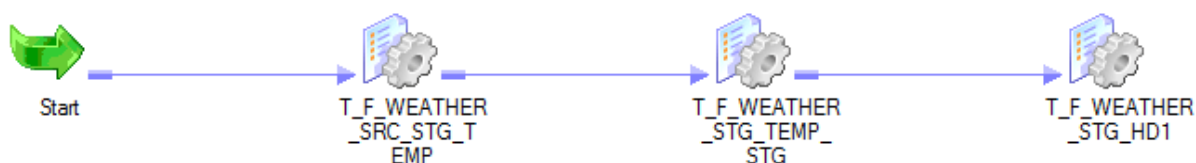


4. F_FLIGHT_STG_HD1

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_F_FLIGHTS, F_FLIGHTS	<p>Ładowanie danych z tabel w STG i HD do tabeli w HD.</p> <p>Generowane jest DWH_ID (max pobierane jest z hurtowni) oraz wyliczana delta między STG i HD.</p> <p>Naliczane są klucze obce do tabel D_AIRPORT oraz D_PLANE_DATA, D_CALENDAR</p> <p>Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatyci.</p>	F_FLIGHTS

6.2.7. WL_F_WEATHER_SRC_HD1

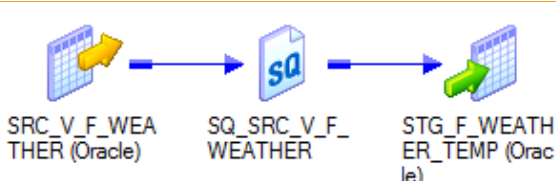
Worklet ładujący dane tabeli F_WEATHER.



Rysunek 6.10 Worklet WL_F_WEATHER_SRC_HD1

1. F_WEATHER_SRC_STG_TEMP

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
SRC_V_F_WEATHER	Ładowanie danych z widoku źródłowego do tabeli tymczasowej w STG. Widok został stworzony do sterowania ilością danych wejściowych. Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	STG_F_WEATHER_TEMP

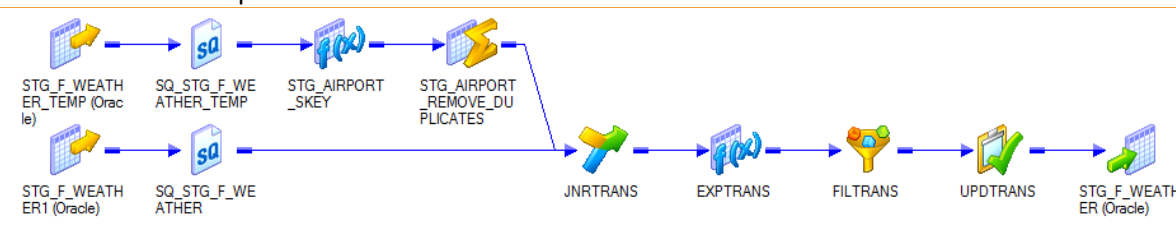


```

graph LR
    A[SRC_V_F_WEATHER (Oracle)] --> B[SQ_SRC_V_F_WEATHER]
    B --> C[STG_F_WEATHER_TEMP (Oracle)]
  
```

2. F_WEATHER_STG_TEMP_STG

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_F_WEATHER_TEMP, STG_F_WEATHER	Ładowanie danych z tabel tymczasowych w STG do tabeli głównej STG. Wyliczany jest SKEY, delta, usuwane duplikaty. Dane są czyszczone oraz transformowane do postaci końcowej. Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	STG_F_WEATHER

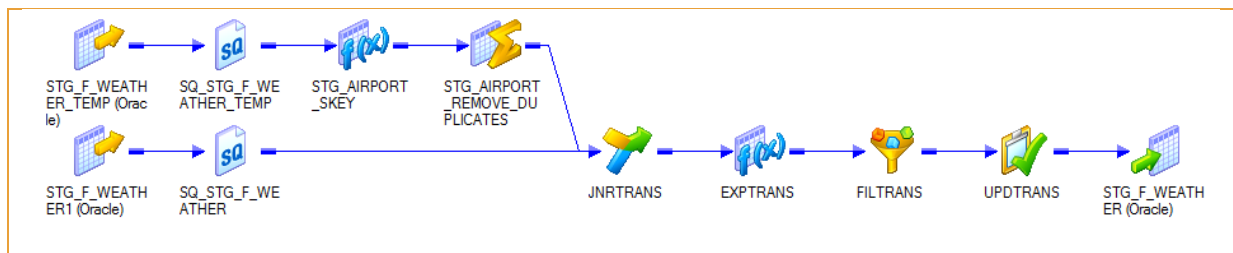


```

graph LR
    A[STG_F_WEATHER_TEMP (Oracle)] --> B[SQ_STG_F_WEATHER_TEMP]
    C[STG_F_WEATHER (Oracle)] --> D[SQ_STG_F_WEATHER]
    B --> E[STG_AIRPORT_SKEY]
    D --> E
    E --> F[STG_AIRPORT_REMOVE_DUPLICATES]
    F --> G[JNRTRANS]
    G --> H[EXPTRANS]
    H --> I[FILTRANS]
    I --> J[UPDTRANS]
    J --> K[STG_F_WEATHER (Oracle)]
  
```

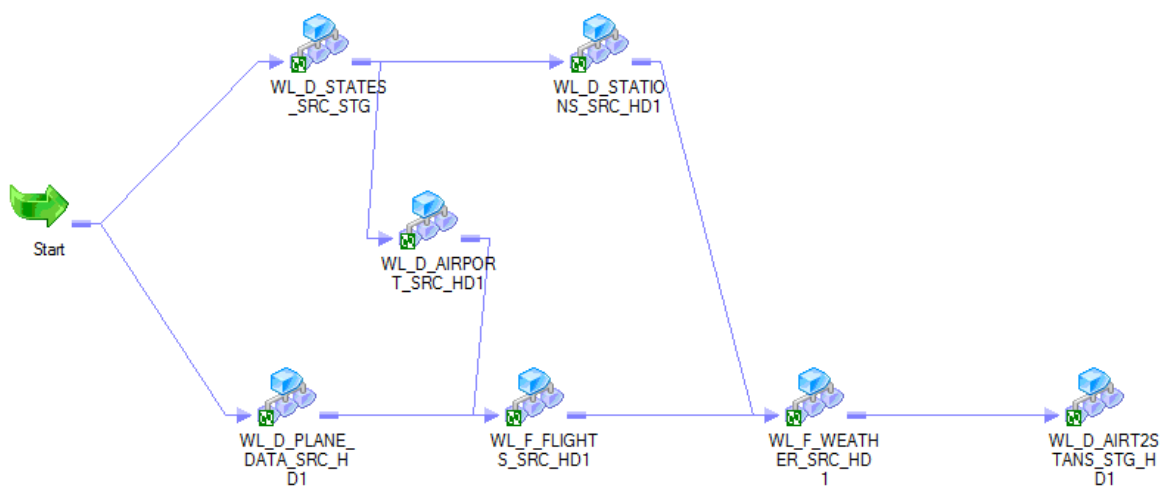
3. F_WEATHER_STG_HD1

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_F_WEATHER_TEMP, STG_F_WEATHER	Ładowanie danych z tabel w STG i HD do tabeli w HD. Generowane jest DWH_ID (max pobierane jest z hurtowni) oraz wyliczana delta między STG i HD. Naliczane są klucze obce do tabeli D_CALENDAR Wszystkie agregacje oraz ładowanie odbywają się za pośrednictwem Informatici.	F_WEATHER



6.1. Workflow WF_ELT

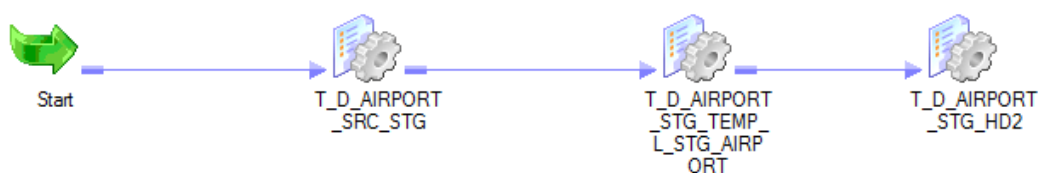
Workflow składa się z 7 workletów odpowiadających ładowaniu 7 tabel ze źródła SOURCE do hurtowni HD2. W następnych podpunktach został sporządzony dokładny opis każdego z nich. Diagram przedstawia zależności pomiędzy poszczególnymi workletami, tj. np. WL_F_FLIGHTS_SRC_HD2 wystartuje dopiero wtedy, kiedy zostaną załadowane 3 inne worklety: WL_D_AIRPORT_SRC_HD2, WL_D_STATES_SRC_STG oraz WL_D_PLANE_DATA_SRC_HD2



Rysunek 6.11 Workflow WF_ELT


6.1.1. WL_D_AIRPORT_SRC_HD2

Worklet ładujący dane tabeli D_AIRPORT.

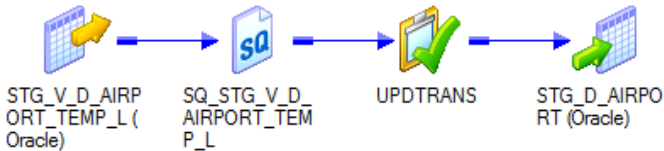


Rysunek 6.12 Worklet WL_D_AIRPORT_SRC_HD2

1. D_AIRPORT_SR->STG_TEMP

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
SRC_D_AIRPORT	Ładowanie danych z tabeli źródłowej do tymczasowej w STG. Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	STG_D_AIRPORT_TEMP
 <pre> graph LR A[SRC_D_AIRPORT (Oracle)] --> B[SQ_SRC_D_AIRPORT] B --> C[STG_D_AIRPORT_TEMP (Oracle)] </pre>		

2. D_AIRPORT_STG_TEMP->STG

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_AIRPORT_TEMP, D_AIRPORT	Ładowanie danych z tabel tymczasowych w STG i HD do tabeli głównej STG. Wyliczany jest SKEY, delta, usuwane duplikaty. Dane są czyszczone oraz transformowane do postaci końcowej. Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	STG_D_AIRPORT
 <pre> graph LR A[STG_V_D_AIRPORT_TEMP_L (Oracle)] --> B[SQ_STG_V_D_AIRPORT_TEMP_L] B --> C[UPDTRANS] C --> D[STG_D_AIRPORT (Oracle)] </pre>		

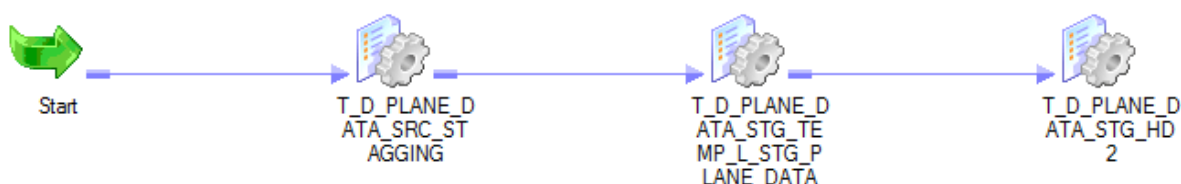
3. D_AIRPORT_STG->HD2

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_AIRPORT	Ładowanie danych z tabel w STG i HD do tabeli w HD. Generowane jest DWH_ID (max pobierane jest z hurtowni) oraz wyliczana delta między STG i HD Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	D_AIRPORT



6.1.2. WL_D_PLANE_DATA_SRC_HD2

Worklet ładujący dane tabeli D_PLANE_DATA.



Rysunek 6.13 Worklet WL_D_PLANE_DATA_SRC_HD2

1. D_PLANE_DATA_SRC->STG_TEMP

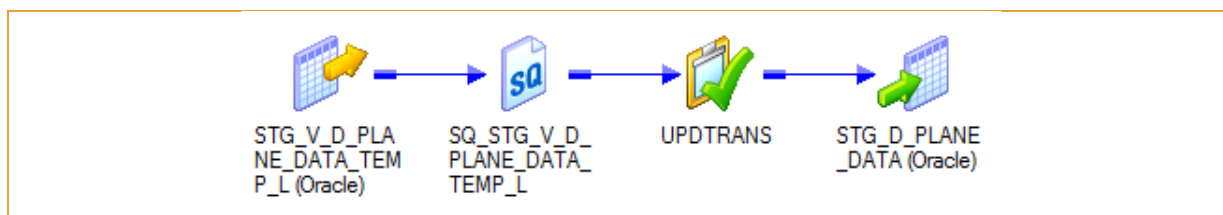
TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
SRC_D_PLANE_DATA	Ładowanie danych z tabeli źródłowej do tymczasowej w STG. Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	STG_D_PLANE_DATA_TEMP


```

graph LR
    A[SRC_D_PLANE_DATA (Oracle)] --> B[SQ_SRC_D_PLANE_DATA]
    B --> C[STG_D_PLANE_DATA_TEMP (Oracle)]
  
```

2. D_PLANE_DATA_STG_TEMP_L->STG_D_PLANE_DATA

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_PLANE_DATA_TEMP	Ładowanie danych z tabel tymczasowych w STG i HD do tabeli głównej STG. Wyliczany jest SKEY, delta, usuwane duplikaty. Dane są czyszczone oraz transformowane do postaci końcowej. Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	STG_D_PLANE_DATA



3. D_PLANE_DATA_STG->HD2

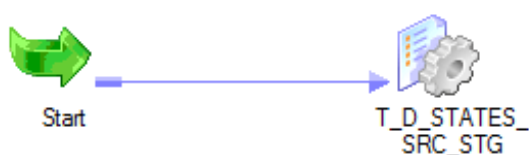
TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_PLANE_DATA, D_PLANE_DATA	Ładowanie danych z tabel w STG i HD do tabeli w HD. Generowane jest DWH_ID (max pobierane jest z hurtowni) oraz wyliczana delta między STG i HD Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	D_PLANE_DATA


```

graph LR
    A[STG_V_D_PLANE_DATA_TEMP_L (Oracle)] --> B[SQ_STG_V_D_PLANE_DATA_TEMP_L1]
    B --> C[D_PLANE_DATA (Oracle)]
  
```

6.1.3. WL_D_STATES_SRC_STG

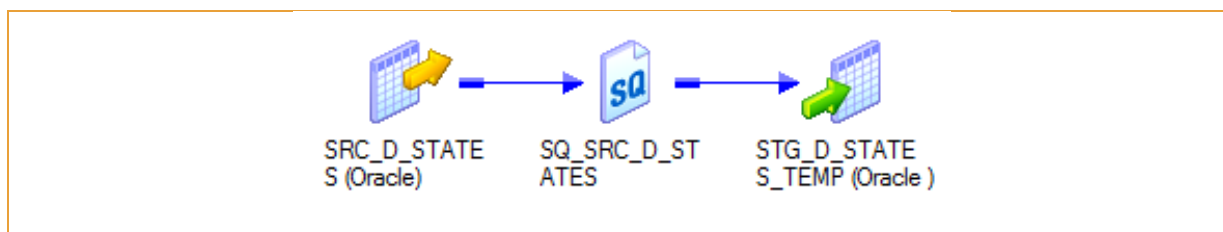
Worklet ładujący dane tabeli D_STATES, ale tylko do STG. Następnie tablica integrowana jest z STG_D_AIRPORT oraz STG_D_STATIONS.



Rysunek 6.14 Worklet WL_D_STATES_SRC_STG

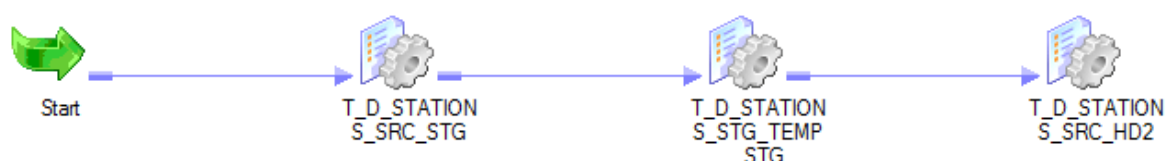
1. D_STATES_SR->STG_TEMP

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
SRC_D_STATES	Ładowanie danych z tabeli źródłowej do tymczasowej w STG. Dane z tej tabeli są integrowane z tabelami STG_D_AIRPORT oraz STG_D_STATIONS Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	STG_D_STATES



6.1.4. WL_D_STATIONS_SRC_HD2

Worklet ładujący dane tabeli D_STATIONS.



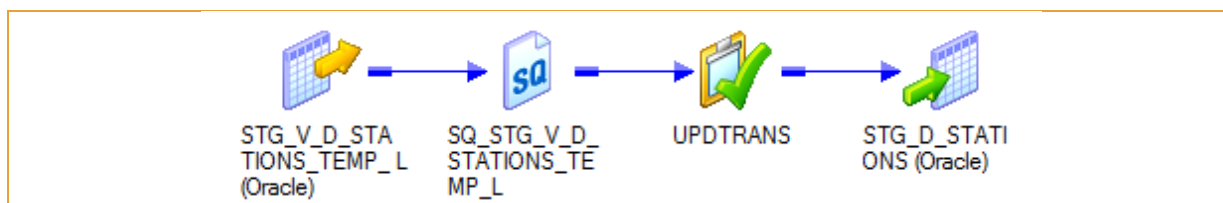
Rysunek 6.15 Worklet WL_D_STATIONS_SRC_HD2

1. D_STATIONS_SR->STG_TEMP

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
SRC_D_STATI ONS	Ładowanie danych z tabeli źródłowej do tymczasowej w STG. Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	STG_D_STATIONS_TEMP
<pre> graph LR A[SRC_D_STATI ONS (Oracle)] --> B[SQ_SRC_D_ST ATIONS] B --> C[STG_D_STATI ONS_TEMP (Oracle)] </pre>		

2. D_STATIONS_STG_TEMP_L->STG_D_STATIONS

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_STATI ONS, D_STATIONS	Ładowanie danych z tabel tymczasowych w STG i HD do tabeli głównej STG. Wyliczany jest SKEY, delta, usuwane duplikaty. Dane są czyszczone oraz transformowane do postaci końcowej. Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	STG_D_STATIONS



3. D_STATIONS_STG->HD2

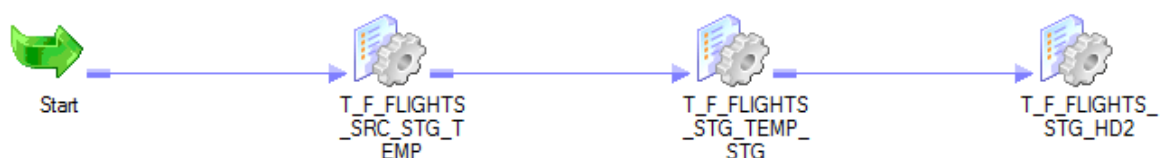
TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_STATIONS	Ładowanie danych z tabel w STG i HD do tabeli w HD. Generowane jest DWH_ID (max pobierane jest z hurtowni) oraz wyliczana delta między STG i HD Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	D_STATIONS


```

graph LR
    A[STG_V_D_STATIONS_L (Oracle)] --> B[SQ_STG_V_D_STATIONS_L]
    B --> C[D_STATIONS (Oracle)]
  
```

6.1.5. WL_F_FLIGHTS_SRC_HD2

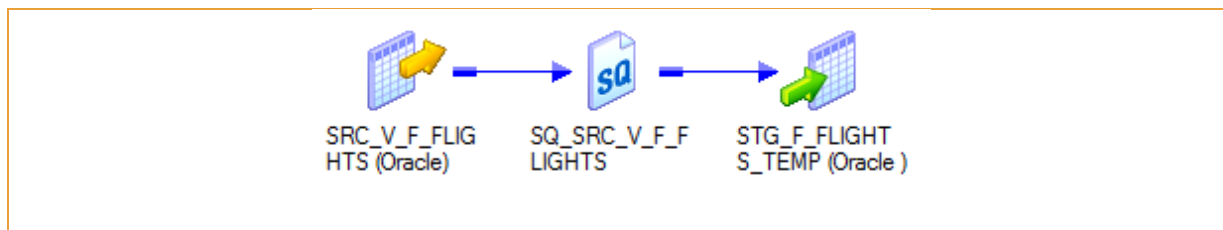
Worklet ładujący dane tabeli F_FLIGHTS.



Rysunek 6.16 Worklet WL_F_FLIGHTS_SRC_HD2

1. F_FLIGHTS_SR->STG_TEMP

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
SRC_F_FLIGHTS	Ładowanie danych z tabeli źródłowej do tymczasowej w STG. Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	STG_F_FLIGHTS_TEMP



2. F_FLIGHTS_STG_TEMP->STG

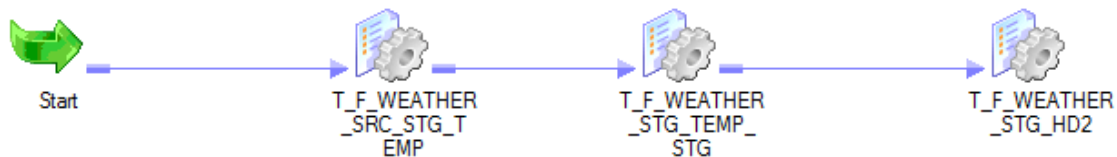
TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_F_FLIGHTS, F_FLIGHTS	Ładowanie danych z tabel tymczasowych w STG i HD do tabeli głównej STG. Wyliczany jest SKEY, delta, usuwane duplikaty. Dane są czyszczone oraz transformowane do postaci końcowej. Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	STG_F_FLIGHTS
<pre> graph LR A[STG_V_F_FLIG HTS_L (Oracle)] --> B[SQ_STG_V_F_F LIGHTS_L] B --> C[UPDTRANS] C --> D[STG_F_FLIGHT S (Oracle)] </pre>		

3. F_FLIGHTS_STG->HD2

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_F_FLIGHTS	Ładowanie danych z tabel w STG i HD do tabeli w HD. Generowane jest DWH_ID (max pobierane jest z hurtowni) oraz wyliczana delta między STG i HD Naliczane są klucze obce do tabel D_AIRPORT oraz D_PLANE_DATA, D_CALENDAR. Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	F_FLIGHTS
<pre> graph LR A[STG_V_F_FLIG HTS_HD_L (Oracle)] --> B[SQ_STG_V_F_F LIGHTS_HD_L] B --> C[F_FLIGHTS (Oracle)] </pre>		

6.1.6. WL_F_WEATHER_SRC_HD2

Worklet ładujący dane tabeli F_WEATHER.



Rysunek 6.17 Worklet WL_F_WEATHER_SRC_HD2

1. F_WEATHER_SR->STG_TEMP

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
SRC_F_WEATHER	Ładowanie danych z tabeli źródłowej do tymczasowej w STG. Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	STG_F_WEATHER_TEMP
<pre> graph LR SRC_V_F_WEATHER[SRC_V_F_WEATHER (Oracle)] --> SQ_SRC_V_F_WEATHER[SQ_SRC_V_F_WEATHER] SQ_SRC_V_F_WEATHER --> STG_F_WEATHER_TEMP[STG_F_WEATHER_TEMP (Oracle)] </pre>		

2. F_WEATHER_STG_TEMP->STG

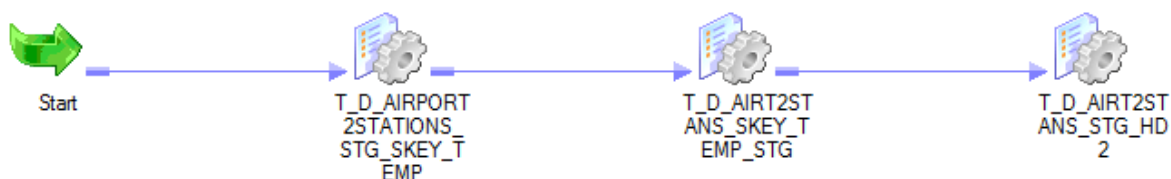
TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_F_WEATHER, F_WEATHER	Ładowanie danych z tabel tymczasowych w STG i HD do tabeli głównej STG. Wyliczany jest SKEY, delta, usuwane duplikaty. Dane są czyszczone oraz transformowane do postaci końcowej. Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.	STG_F_WEATHER
<pre> graph LR STG_V_F_WEATHER_L[STG_V_F_WEATHER_L (Oracle)] --> SQ_STG_V_F_WEATHER_L[SQ_STG_V_F_WEATHER_L] SQ_STG_V_F_WEATHER_L --> UPDTRANS[UPDTRANS] UPDTRANS --> STG_F_WEATHER[STG_F_WEATHER (Oracle)] </pre>		

3. F_WEATHER_STG->HD2

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_F_WEATHER	<p>Ładowanie danych z tabel w STG i HD do tabeli w HD.</p> <p>Generowane jest DWH_ID (max pobierane jest z hurtowni) oraz wyliczana delta między STG i HD.</p> <p>Naliczane są klucze obce do tabeli D_CALENDAR.</p> <p>Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.</p>	F_WEATHER
<pre> graph LR A[STG_V_F_WEATHER_HD_L (O)] --> B[SQ_STG_V_F_WEATHER_HD] B --> C[F_WEATHER (Oracle)] </pre>		

6.1.7. WL_D_AIRT2STATIONS_STG_HD2

Worklet ładujący dane tabeli D_AIRPORT2STATIONS.



Rysunek 6.18 Worklet WL_D_AIRT2STATIONS_STG_HD2

1. D_AIRT2STANS_SKEY_TEMP->STG_TEMP

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
D_AIRPORT, D_STATIONS	<p>Ładowanie danych z tabel z HD do tymczasowej w STG.</p> <p>Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.</p>	STG_D_AIRT2STANS_SKEY_TEMP
<pre> graph LR A[STG_D_V_AIRT2STANS_SKEY_TEMP_L (Oracle)] --> B[SQ_STG_D_V_AIRT2STANS_SKEY_TEMP_L] B --> C[STG_D_AIRT2STANS_SKEY_TEMP (Oracle)] </pre>		

2. D_AIRT2STATNS_STG_TEMP->STG

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_AIRT2STANS_SKEY_TEMP	<p>Ładowanie danych z tabel z STG do tabeli głównej w STG.</p> <p>Dane są transformowane do postaci końcowej. Wyliczany jest MIN dystans pomiędzy lotniskami i stacjami pogodowymi, do tabeli wstawiane są klucze związane tylko z najkrótszą drogą między punktami.</p> <p>Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.</p>	STG_D_AIRPORT2STATIONS
<pre> graph LR A[STG_D_V_AIRT2STANS_L (Oracle)] --> B[SQ_STG_D_V_AIRT2STANS_L] B --> C[UPDTRANS] C --> D[STG_D_AIRPORT2STATIONS (Oracle)] </pre>		

3. D_AIRT2STANS_STG->HD2

TABELA ŹRÓDŁOWA	OPERACJE	TABELA DOCELOWA
STG_D_AIRPORT2STATIONS, D_AIRPORT2STATIONS	<p>Ładowanie danych z tabel z STG do tabeli w HD.</p> <p>Generowane jest DWH_ID (max pobierane jest z hurtowni) oraz wyliczana delta między STG i HD</p> <p>Wszystkie agregacje odbywają się w bazie danych, ładowanie odbywa się za pośrednictwem Informatici.</p>	D_AIRPORT2STATIONS
<pre> graph LR A[STG_D_V_AIRT2STANS_HD_L (Oracle)] --> B[SQ_STG_D_V_AIRT2STANS_HD_L] B --> C[UPDTRANS] C --> D[D_AIRPORT2STATIONS (Oracle)] </pre>		

7. Testy

7.1. Procedura testowa

Procedura testowa składa się z 4 etapów, w których dane były ładowane stopniowo, ze względu na wartości atrybutów YEAR w SRC_F_FLIGHT oraz SRC_F_WEATHER. Tablice wymiarów zostały załadowane w całości przy pierwszym podejściu, jednakże w celu wykrycia usuniętych rekordów, zostały ponownie załadowane na kolejnych etapach do tablic tymczasowych w STG gdzie wyliczona została delta. To oznacza, że przy każdym ładowaniu zostały załadowane do tablic tymczasowych wszystkie dostępne dane i dopiero później odrzucone przy obliczaniu delty. W ten sposób do minimum ograniczamy obciążenie źródła, ponieważ operacja Select jest wielokrotnie szybsza od wykonywania agregacji bezpośrednio na źródle.

7.2. Etapy testowe

I ładowanie

Ładowanie danych z roku 1987

II ładowanie

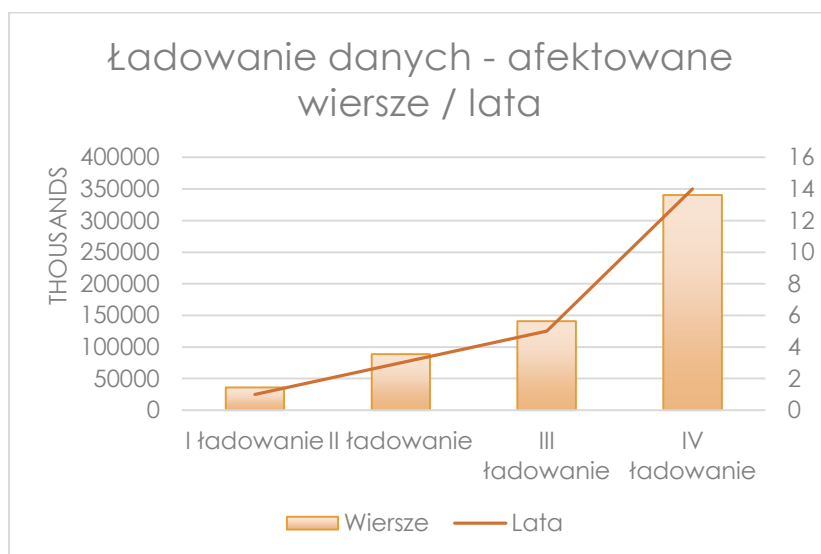
Ładowanie danych z lat 1987-1989

III ładowanie

Ładowanie danych z lat 1987-1992

IV ładowanie

Ładowanie danych z lat 1987-2000



Rysunek 7.1 Ładowanie danych - afektowane wiersze / lata

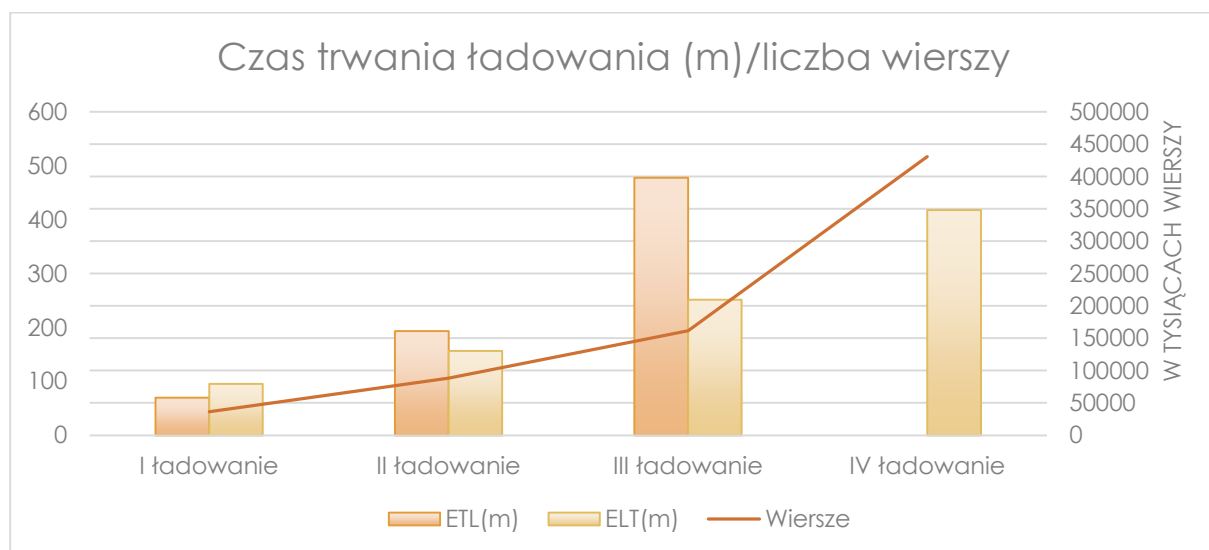
7.3. Wyniki

Testy dla ETL oraz ELT zostały przeprowadzone niezależnie, by zminimalizować wpływ jednych przepływów na drugie. Wykonano 8 przepływów danych, 7 zakończyło się sukcesem.

Tabela poniżej przedstawia zestawienie czasów oraz ilości pobranych wierszy ze źródła. Wynika z tego, iż baza danych znacznie lepiej sobie radzi przy większych wolumenach danych. W przypadku ładowania ETL widać, że w pewnym momencie ładowanie przestaje być wydajne, co ma związek z przepełnieniem się pamięci operacyjnej oraz faktem, iż środowisko oparte jest o maszyny wirtualne.

Etap	Rok	Wiersze	ETL (m)	ELT (m)
I	1987	36283213	69.6	95.27
II	1987-1989	88551676	193.3	156.28
III	1987-1992	161402440	477.43	251.62
IV	1987-2000	430673661	-	417.75

Tabela 7.1 Podsumowanie przeprowadzonych testów



Rysunek 7.2 Czas trwania ładowania (m)/ ilość wierszy

Poniżej zostały opisane wyniki detaliczne pomiarów. Tabele zawierają ilości wierszy, które zostały pobrane ze źródła, nie odzwierciedla to jednak faktycznej ilości przetworzonych wierszy, ponieważ na poszczególnych etapach ładowania wiersze są agregowane wielokrotnie, usuwane są duplikaty, itd. W dwóch pozostałych kolumnach znajdują się pomiary czasowe wyrażone w minutach. Są to sumy czasów przypadających na wszystkie etapy ładowania danych (pobranie danych ze źródła, transformacja w STG i końcowe ładowanie do HD).

7.3.1. Etap I – dane za rok 1987

Tablica	Wiersze	WF_ETL (m)	WF_ELT (m)
WL_D_AIRT2STANS	98130	65.58	14.38
WL_D_PLANE_DATA	20128	0.11	0.09
WL_D_AIRPORT	13504	0.11	0.09
WL_D_STATIONS	274206	0.22	0.03
WL_F_FLIGHTS	12999894	2.47	45.13
WL_D_STATES	73	0.03	0.03
WL_F_WEATHER	22877278	1.08	35.52
Total	36283213	69.6	95.27

Tabela 7.2 Etap I – dane za rok 1987

7.3.2. Etap II – dane za lata 1987-1989

Tablica	Wiersze	WF_ETL (m)	WF_ELT (m)
WL_D_AIRT2STANS	98130	94.38	18.01
WL_D_PLANE_DATA	20128	0.13	0.09
WL_D_AIRPORT	13504	0.11	0.1
WL_D_STATIONS	274206	0.22	0.05
WL_F_FLIGHTS	28992235	66.23	75.48
WL_D_STATES	73	0.04	0.03
WL_F_WEATHER	59153400	32.19	62.52
Total	88551676	193.3	156.28

Tabela 7.3 Etap II – dane za lata 1987-1989

7.3.3. Etap III – dane za lata 1987-1992

Tablica	Wiersze	WF_ETL (m)	WF_ELT (m)
WL_D_AIRT2STANS	98130	174.81	33.08
WL_D_PLANE_DATA	20128	0.13	0.12
WL_D_AIRPORT	13504	0.15	0.17
WL_D_STATIONS	274206	0.22	0.07
WL_F_FLIGHTS	63854210	164.1	182.62
WL_D_STATES	73	0.04	0.04
WL_F_WEATHER	97142189	137.98	35.52
Total	161402440	477.43	251.62

Tabela 7.4 Etap III – dane za lata 1987-1992

7.3.4. Etap IV – dane za lata 1987-2000

Tablica	Wiersze	WF_ETL (m)	WF_ELT (m)
WL_D_AIRT2STANS	98130	274.81	37.07
WL_D_PLANE_DATA	20128	0.13	0.12
WL_D_AIRPORT	13504	0.15	0.17
WL_D_STATIONS	274206	0.22	0.07
WL_F_FLIGHTS	173765431	-	278.13
WL_D_STATES	73	0.04	0.04
WL_F_WEATHER	256502189	411.3	102.15
Total	430673661	-	417.75

Tabela 7.5 Etap IV – dane za lata 1987-2000

8. Wnioski i podsumowanie

Z testów wynika jednoznacznie, iż największy wpływ na przebieg doświadczeń miały dyski twarde oraz fakt wykorzystania maszyn wirtualnych.

Trudno jednak stwierdzić, które rozwiązanie, ETL lub ELT, jest lepsze oraz bardziej wydajne. Warto stosować metody hybrydowe, wykorzystywać najmocniejsze strony każdego z podejść. Wyraźnie ELT radzi sobie lepiej przy łączeniu dużych tabel, natomiast Informatica poradziła sobie lepiej przy transformacji atrybutów, co więcej posiada bardzo szybkie lookupy.

Warto użyć narzędzia ETL do integracji danych z wielu różnych źródeł, ale wszelkie transformacje należy pozostawić RDBMS. Informatica posiada wysokowydajne sterowniki do wielu popularnych silników bazodanowych, potrafi parsować pliki CSV, XML oraz pobierać dane wprost z webserwisów. Niestety sam interfejs jest bardzo skomplikowany oraz cechuje się dużą ilością wad i błędów. Bardzo trudno debugować kod skryptu. W wielu miejscach nie występuje podpowiadanie kodu, kompilator nie znajduje błędów (jedynym objawem ich występowania jest brak załadowania danych przez workflow).

Silnik Oracle nie zgłaszał problemów nawet przy skrajnych obciążeniach. Informatica jest bardzo wrażliwa na ilość pamięci operacyjnej w systemie, nie wytrzymywała rywalizacji o zasoby z silnikiem bazodanowym (często się „zawieszała”).

W środowisku produkcyjnym należy instalować narzędzia ETL, repozytoria oraz docelowe bazy danych na oddzielnych serwerach, tak by nie rywalizowały ze sobą o moc obliczeniową. Tylko w takim przypadku narzędzia ETL będą wydajnie wspomagać procesy ładowania danych.

9. Literatura i źródła danych

Alkis Simitsis, Panos Vassiliadis, Umeshwar Dayal, Anastasios Karagiannis, Vasiliki Tziouvara. *Benchmarking ETL Workflows*. University of Ioannina.

Panos Vassiliadis, Alkis Simitsis, Eftychia Baikousi. *A Taxonomy of ETL Activities*.

Anastasios Karagiannis, Panos Vassiliadis, Alkis Simitsis. *Macro-level Scheduling of ETL Workflows*.

Kozielski S., Małysiak B., Kasprowski P., Mrozek D.. *Bazy Danych: Nowe Technologie*. 2007

Ploug. <http://www.ploug.org.pl/>

Materiały udostępnione w ramach studiów Politechniki Poznańskiej,

http://tpd.cs.put.poznan.pl/studia-podyplomowe/sp_hd/

Dokumentacja ORACLE, <http://www.oracle.com/index.html>

Dokumentacja INFORMATICA, <https://community.informatica.com/docs>

10. Spis rysunków

Rysunek 2.1 Model architektury ETL.....	4
Rysunek 2.2 Model architektury ELT.....	4
Rysunek 3.1 Architektura oraz model przepływu	7
Rysunek 4.1 Schemat bazy SOURCE	14
Rysunek 5.1 Schemat hurtowni danych HD	19
Rysunek 6.1 Model ładowania ELT.....	21
Rysunek 6.2 Model ładowania ETL.....	22
Rysunek 6.3 Workflow WF_ETL.....	22
Rysunek 6.4 Worklet WL_D_AIRPORT_SRC_HD1.....	23
Rysunek 6.5 Worklet WL_D_AIRT2STANS_STG_HD1.....	24
Rysunek 6.6 Worklet WL_D_PLANE_DATA_SRC_HD1	26
Rysunek 6.7 Worklet WL_D_STATES_SRC_STG.....	27
Rysunek 6.8 Worklet WL_D_STATIONS_SRC_HD1.....	28
Rysunek 6.9 Worklet WL_F_FLIGHTS_SRC_HD1	30
Rysunek 6.10 Worklet WL_F_WEATHER_SRC_HD1	31
Rysunek 6.11 Workflow WF_ELT.....	33
Rysunek 6.12 Worklet WL_D_AIRPORT_SRC_HD2	33
Rysunek 6.13 Worklet WL_D_PLANE_DATA_SRC_HD2	35
Rysunek 6.14 Worklet WL_D_STATES_SRC_STG.....	36
Rysunek 6.15 Worklet WL_D_STATIONS_SRC_HD2.....	37
Rysunek 6.16 Worklet WL_F_FLIGHTS_SRC_HD2	38
Rysunek 6.17 Worklet WL_F_WEATHER_SRC_HD2	40
Rysunek 6.18 Worklet WL_D_AIRT2STATIONS_STG_HD2	41
Rysunek 7.1 Ładowanie danych - afektowane wiersze / lata.....	43
Rysunek 7.2 Czas trwania ładowania (m)/ ilość wierszy	44

11. Spis tabel

Tabela 3.1 Specyfikacja serwera głównego – Source.....	7
Tabela 3.2 Specyfikacja serwera wirtualnego – ETL REP.....	8
Tabela 3.3 Specyfikacja serwera wirtualnego – HD+ETL REP.....	8
Tabela 4.1 Struktura tabeli SRC_F_WEATHER.....	10
Tabela 4.2 Struktura tabeli SRC_F_FLIGHTS.....	11
Tabela 4.3 Struktura tabeli SRC_D_AIRPORT	11
Tabela 4.4 Struktura tabeli SRC_D_PLANE_DATA	12
Tabela 4.5 Struktura tabeli SRC_D_STATES.....	12
Tabela 4.6 Struktura tabeli SRC_D_STATIONS	12
Tabela 4.7 Indeksy	13
Tabela 4.8 Charakterystyka tabel.....	13
Tabela 5.1 Struktura tabeli F_WEATHER.....	15
Tabela 5.2 Struktura tabeli D_STATIONS.....	15
Tabela 5.3 Struktura tabeli D_CALENDAR	16
Tabela 5.4 Struktura tabeli F_FLIGHTS	17
Tabela 5.5 Struktura tabeli D_AIRPORT	18
Tabela 5.6 Struktura tabeli D_AIRPORT2STATIONS	18
Tabela 5.7 Struktura tabeli D_PLANE_DATA	18
Tabela 7.1 Podsumowanie przeprowadzonych testów	44
Tabela 7.2 Etap I – dane za rok 1987.....	45
Tabela 7.3 Etap II – dane za lata 1987-1989.....	45
Tabela 7.4 Etap III – dane za lata 1987-1992.....	45
Tabela 7.5 Etap IV – dane za lata 1987-2000	46

12. Załączniki

12.1. Skrypty SQL

12.1.1. HD

- HD1_HD2_Calendar.sql
- HD1_HD2_Tables+FK.sql

12.1.2. Staging

- ELT_STAGGING_D_AIRPORT_DATA.sql
- ELT_STAGGING_D_AIRPORT2STATIONS.sql
- ELT_STAGGING_D_PLANE_DATA.sql
- ELT_STAGGING_D_STATES.sql
- ELT_STAGGING_D_STATIONS.sql
- ELT_STAGGING_F_FLIGHTS.sql
- ELT_STAGGING_F_WEATHER.sql
- ETL_STAGGING_D_AIRPORT_DATA.sql
- ETL_STAGGING_D_AIRPORT2STATIONS.sql
- ETL_STAGGING_D_PLANE_DATA.sql
- ETL_STAGGING_D_STATES.sql
- ETL_STAGGING_D_STATIONS.sql
- ETL_STAGGING_F_FLIGHTS.sql
- ETL_STAGGING_F_WEATHER.sql
- ELT_STAGGING_Indexes.sql

12.1.3. Source

- SourceIndexes.sql
- SourceTables.sql