Wojskowa Akademia Techniczna

Hurtownie Danych

Sprawozdanie z projektu

Autorzy: Grzegorz Bancerz, Tomasz Bandurski
Grupa: I6B2S1
Prowadzący: dr inż. Marcin Mazurek

Treść zadania

Na podstawie danych źródłowych zawierających o dane transporcie lotniczym w USA zbudować hurtownię danych.

Kamienie milowe:

1. Model wymiarowy danych
2. Instalacja środowiska deweloperskiego
3. Model bazy danych repozytorium głównego, model obszaru Stage, załadowane dane do Stage
4. Proces ładujący do repozytorium głównego, załadowane dane do repozytorium głównego
5. Baza Analysis Services (wielowymiarowe kostki danych)
6. Raportowanie ad-hoc w narzędziu Power BI.

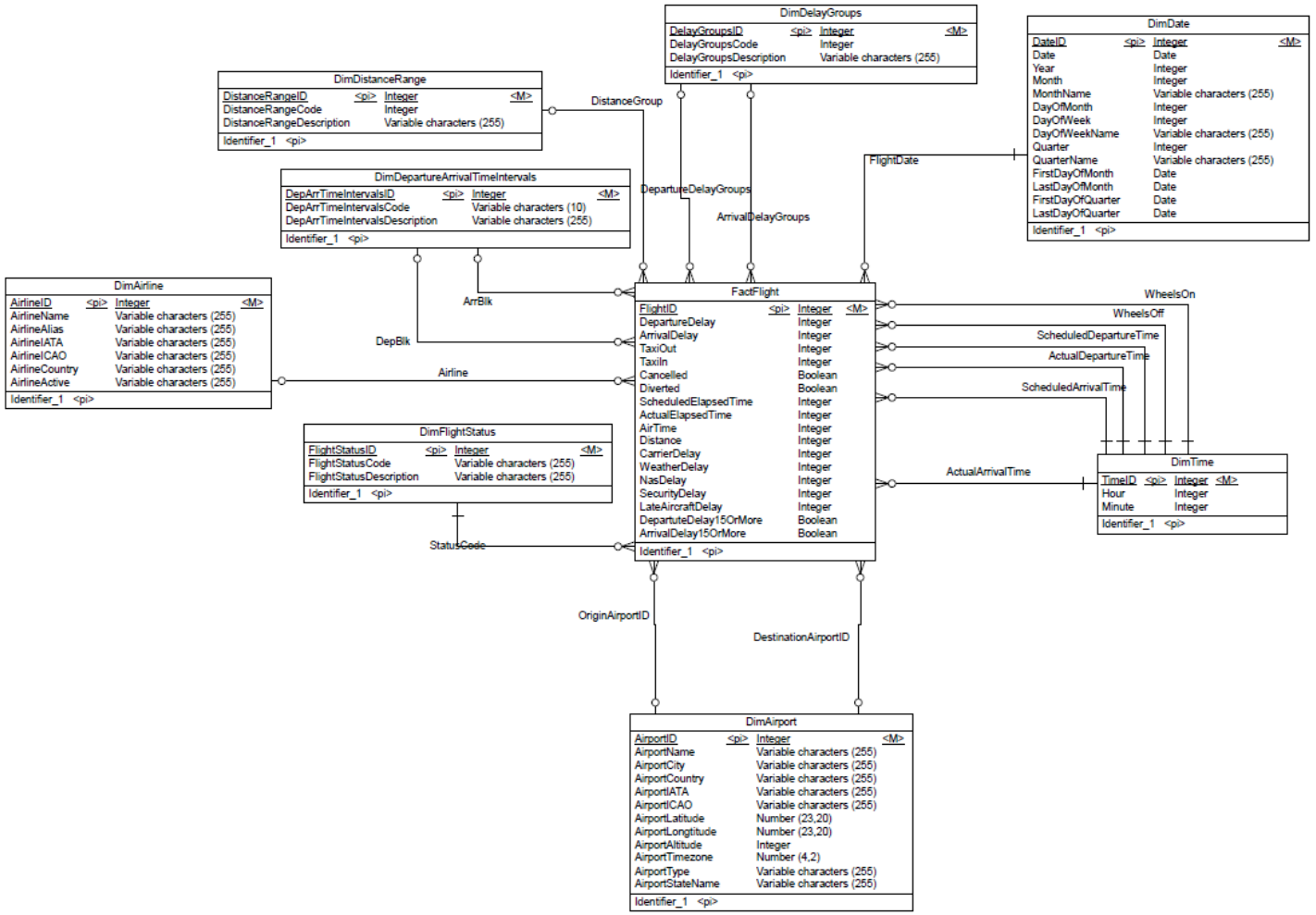
Źródła danych

https://transtats.bts.gov/DL_SelectFields.asp?Table_ID=236

https://openflights.org/data.html

Modele danych

Konceptualny model hurtowni danych



Fizyczny model hurtowni danych

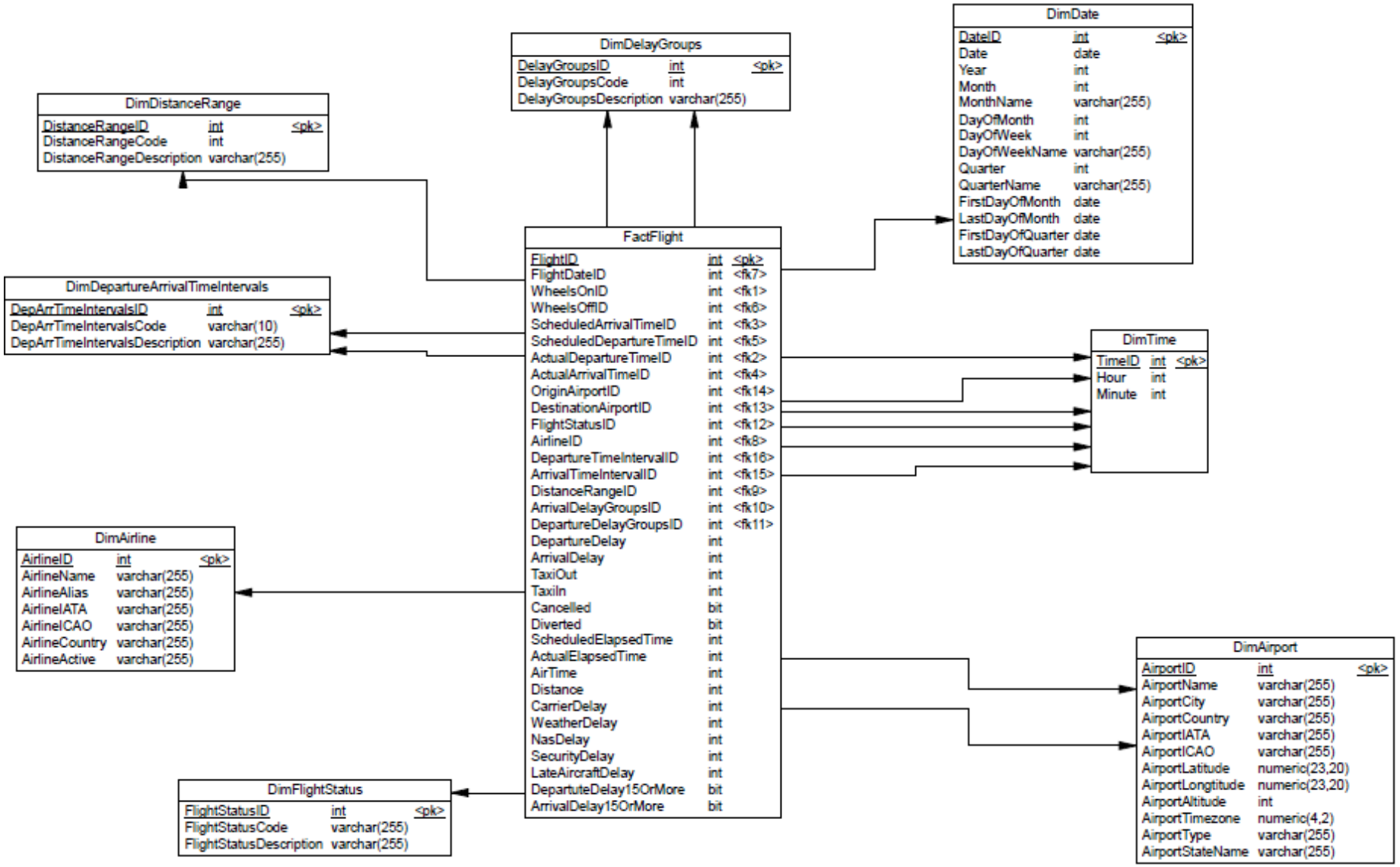


Tabela faktów (za fakt przyjęliśmy pojedynczy lot)

Znaczenie atrybutów

- **DepartureDelay** (Opóźnienie odlotu) [min] - Różnica w minutach między zaplanowanym a faktycznym czasem odlotu. Odloty przed czasem posiadają liczby ujemne.
- **ArrivalDelay** (Opóźnienie przylotu) [min] - Różnica w minutach między zaplanowanym a rzeczywistym czasem przylotu. Przyloty przed czasem posiadają liczby ujemne.
- **TaxiOut** (Czas kołowania przed wylotem) [min] - okres pomiędzy czasem, w którym samolot opuszcza pozycję parkowania a czasem, w którym podnosi się z pasa startowego
- **TaxiIn** (Czas kołowania po przylocie) [min] - okres pomiędzy czasem lądowania samolotu a zatrzymaniem samolotu w pozycji parkowania
- **Cancelled** (Anulowany) - wartość 1 gdy lot był anulowany, 0 w przeciwnym przypadku
- **Diverted** (Zawrócony) - wartość 1 gdy lot był zawrócony, 0 w przeciwnym przypadku
- **ScheduledElapsedTime** (Zaplanowany czas lotu) [min]
- **ActualElapsedTime** (Rzeczywisty czas lotu) [min]
- **AirTime** (Czas lotu) [min]
- **Distance** (Odległość) [mila] - Odległość między lotniskami
- **CarrierDelay** (Opóźnienie spowodowane przez przewoźnika) [min]
- **WeatherDelay** (Opóźnienie spowodowane przez pogodę) [min]
- **NasDelay** (Opóźnienie spowodowane przez National Air System) [min]
- **SecurityDelay** (Opóźnienie spowodowane względami bezpieczeństwa) [min]
- **LateAircraftDelay** (Opóźnienie spowodowane przez inny spóźniony samolot) [min]
- **DepartuteDelay15OrMore** (Opóźnienie odlotu) - wartość 1 gdy opóźnienie odlotu wynosi 15 minut lub więcej, 0 w przeciwnym przypadku
- **ArrivalDelay15OrMore** (Opóźnienie przylotu) - wartość 1 gdy opóźnienie przylotu wynosi 15 minut lub więcej, 0 w przeciwnym przypadku

Tabele wymiarów

- Airline
- Airport
- Date
- DistanceRange
- DepartureArrivalTimeIntervals
- DelayGroups
- FlightStatus
- Time

Utworzone miary

- Fact Flight Count - liczba lotów
- Maximum Arrival Delay – maksymalne opóźnienie przylotu
- Maximum Departure Delay – maksymalne opóźnienie odlotu
- Departure Delay Non Negative - różnica w minutach między zaplanowanym a faktycznym czasem odlotu bez liczb ujemych
- Arrival Delay Non Negative - Różnica w minutach między zaplanowanym a rzeczywistym czasem przylotu bez liczb ujemych
- Flights On Time - liczba lotów na czas
- Punctuality - udział lotów zrealizowanych w czasie do wszystkich (zrealizowane w czasie oznacza, że ich opóźnienie przylotu było poniżej 15 minut)
- Avarage Arrival Delay- średnie opóźnienie przylotu
- Avarage Departure Delay- średnie opóźnienie odlotu
- Delays Sum - całkowita suma opóźnień przylotów i odlotów

Wygenerowane skrypty

- **SkryptStageGenrMSSQL** – skrypt tworzący tabele dla bazy danych obszaru Stage SQL Server
- **SkryptGenrBdMsSQL** – skrypt tworzący tabele dla hurtowni danych SQL Server

Rozwiązania

- **SSISwarehouse** – w pakietach DimAirline, DimAirport, DimDate, DimDelayGroups, DimDepArrBlk, DimDistance, DimFlightStatus, DimTime zasilane są tabele wymiarów. W pakiecie StageFixed zasilane są lookup tables i tabele wymagane do zasilenia wymiarów. Pakiet StageFacts odpowiada za ładownia faktów do Stage. W tym pakiecie dodajemy również dodatkową kolumnę o nazwie FlightStatus, w której będą zawarte kody statusów lotów:
  - DV – Diverted – lot zawrócony
  - DD – Delayed – lot opóźniony, czyli taki, którego opóźnienie przylotu wynosiło powyżej 15 minut
  - OT – On Time – lot zrealizowany w czasie, czyli taki, którego opóźnienie przylotu wynosiło poniżej 15 minut
  - CA, CB, CC, CD – lot odwołany z danej przyczyny (CA - Cancelled due Carrier, CB - Cancelled due Weather, CC - Cancelled due National Air System, CD - Cancelled due Security)

  W pakiecie FactFlight ładowane są dane faktów do repozytorium głównego.

  Za przyrostowe ładowanie danych odpowiada pakiet IncLoadingOfFacts. Jest to połączenie pakietów StageFacts i FactFlight mające na celu uproszczenie uruchamiania procesu przyrostowego ładowania danych.

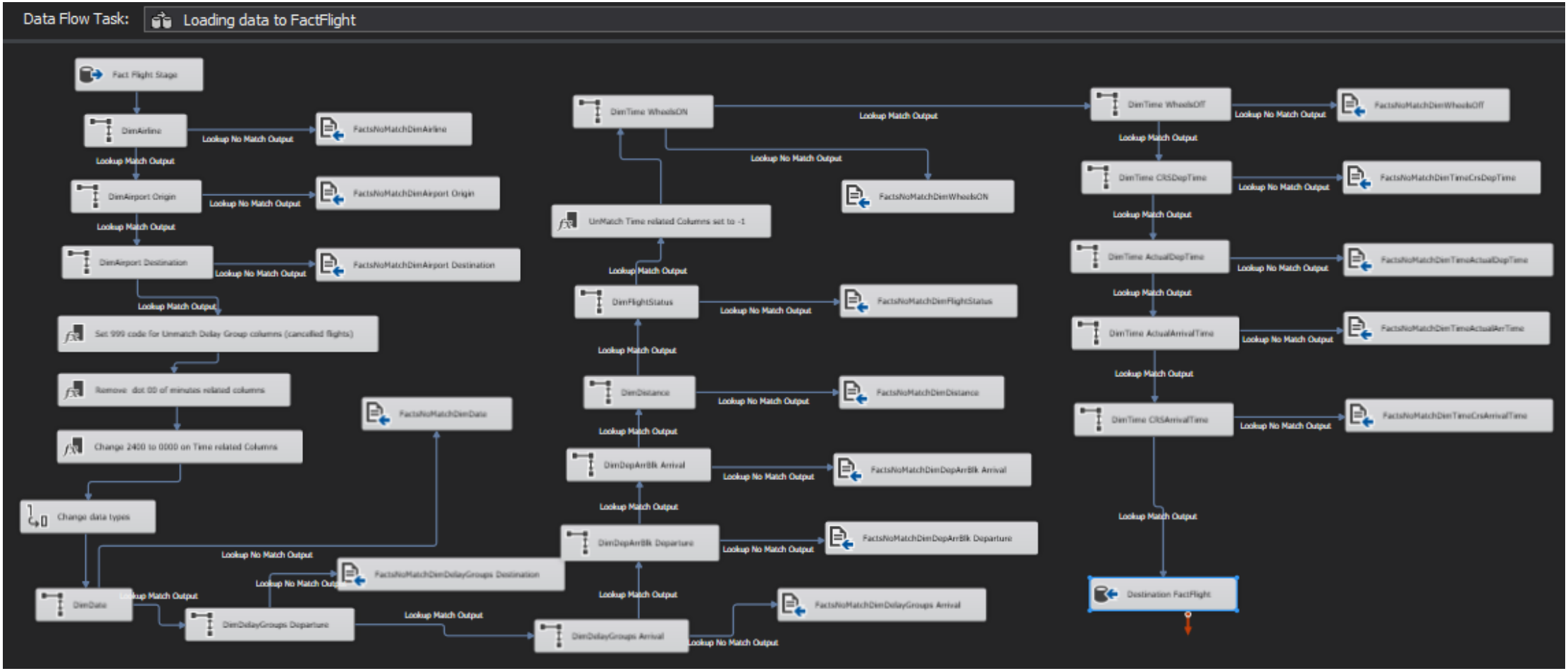
  Założenia ETL:

  - Na NULL zamieniane są puste wartości miar (na przykład opóźnienia spowodowane warunkami atmosferycznymi) i braki danych w wymiarach (DimAirline i DimAirport)
  - W przypadku niedopasowania wartości klucza obcego przyporządkowany jest następujący klucz:

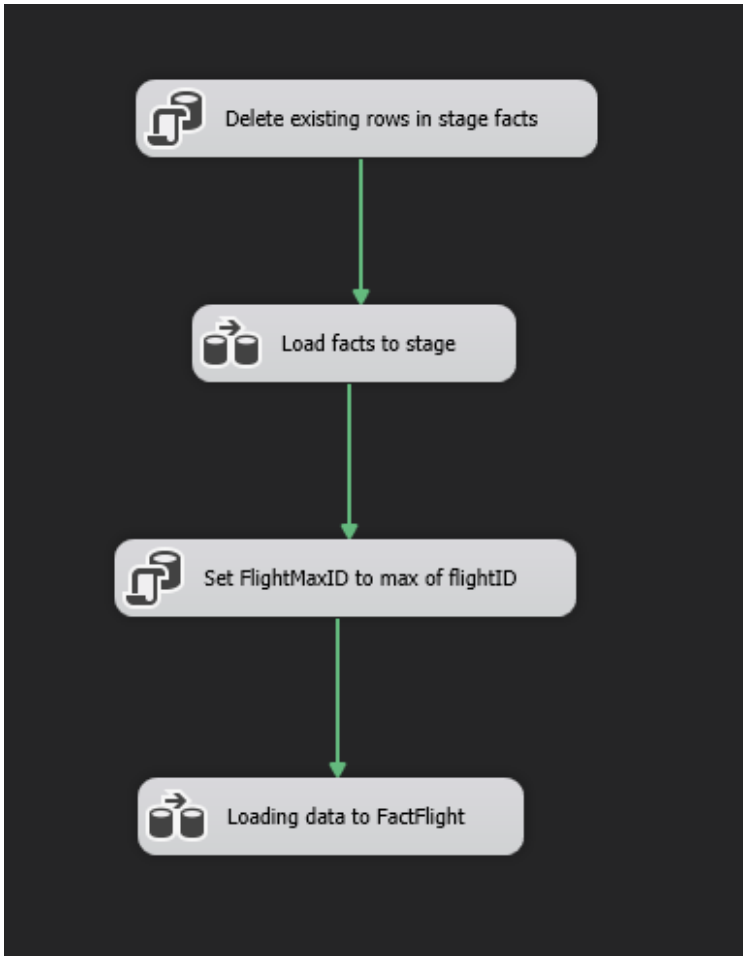
- Dla braku czasu wylotu, przylotu itd. Wartość klucza obcego w DimTime -1
- Dla grup opóźnień DimDelayGroups -1 (oznacza loty anulowane)

Dodatkowe informacje o procesie ładowania faktów:

- Task „Set FlightMaxID to max of flightID” – ustawienie zmiennej na maksimum FlightID z tabeli faktów (do przypisywania wartości klucza w przypadku doładowywania danych)
- Każdy task typu Lookup posiada plik na ewentualne niedopasowane rekordy (w naszym przypadku nie było takich rekordów dla danych z roku 2008)
- Task „Set 999 code for Unmatch Delay Group columns (cancelled flights)” – ustawienie kodu 999 dla niedopasowanych rekordów wymiaru DimDelayGroups (Arrival i Departure)
- Task „Remove dot 00 of minutes related columns” – usunięcie „.00” z odpowiednich kolumn w celu zmiany ich typu
- Task „Change 2400 to 0000 on Time related Columns” – zamiana godziny 24:00 na 00:00 dla odpowiednich kolumn
- Task „UnMatch Time related Columns set to -1” - przypisanie -1 dla pustych czasów kluczy obcych

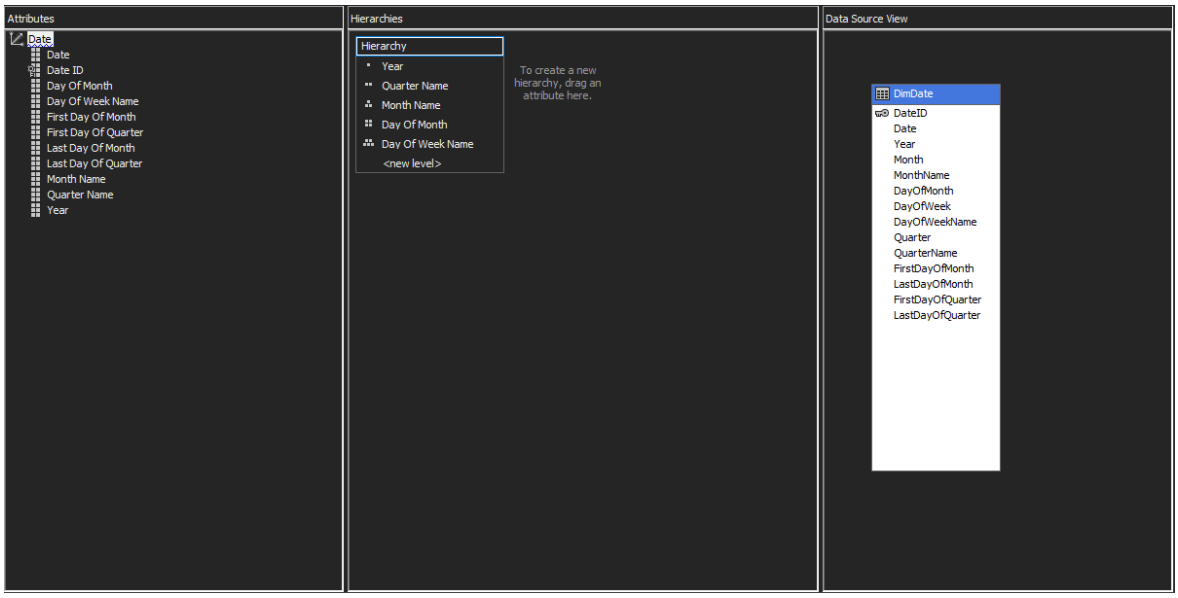


Rys. 1 Task "Loading data to FactFlight"



Rys. 2 Pakiet IncLoadingOfFacts- przyrostowe ładowanie faktów

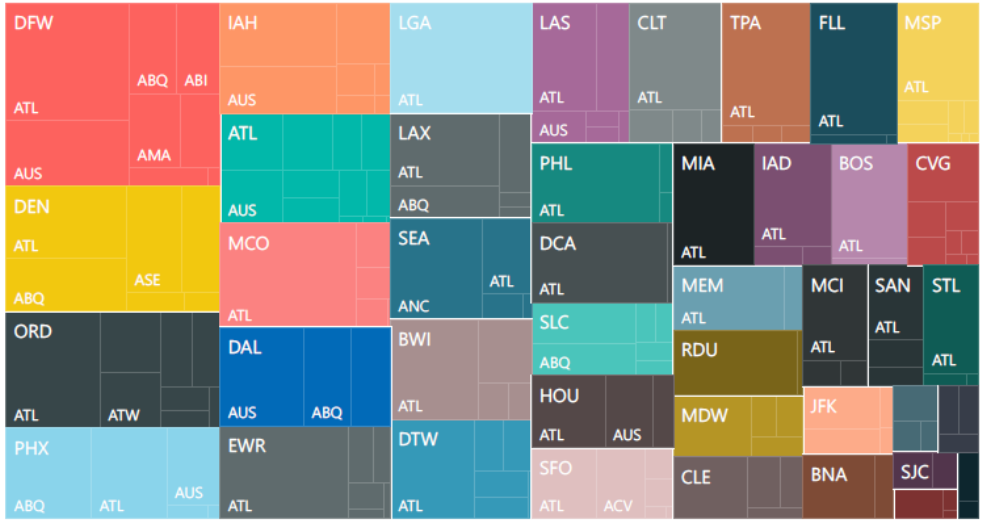
- **SSASFlightsCube** – rozwiązanie zawierające projekt wielowymiarowej kostki danych



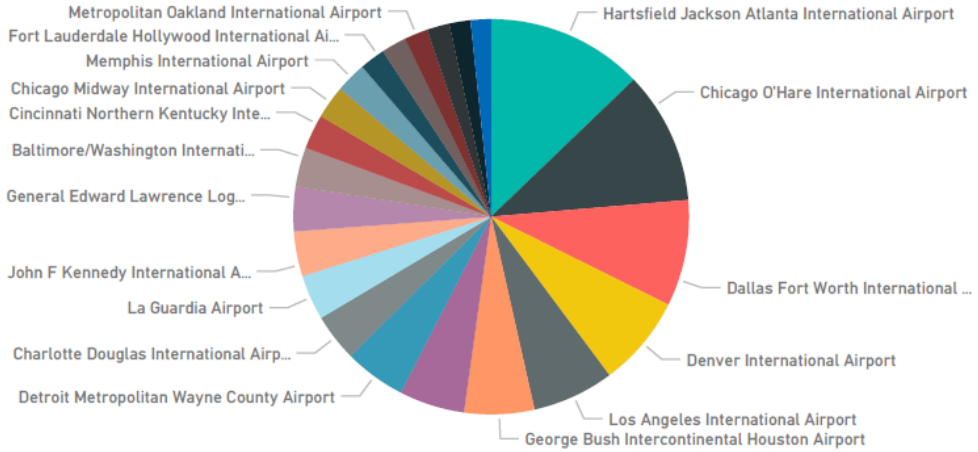
Rys.3 Utworzona hierarchia dla wymiaru DimDate

Przykładowe raporty

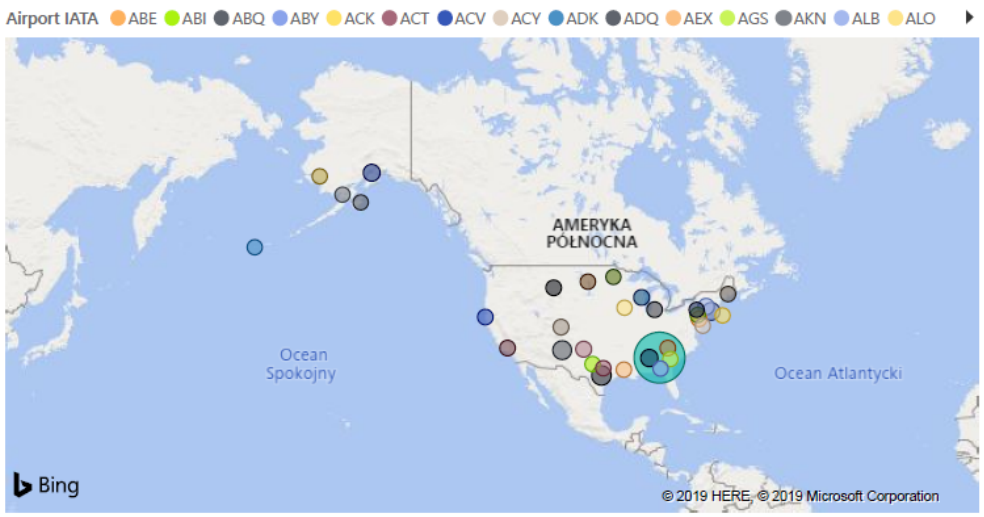
Fact Flight Count wg Airport IATA i Airport IATA



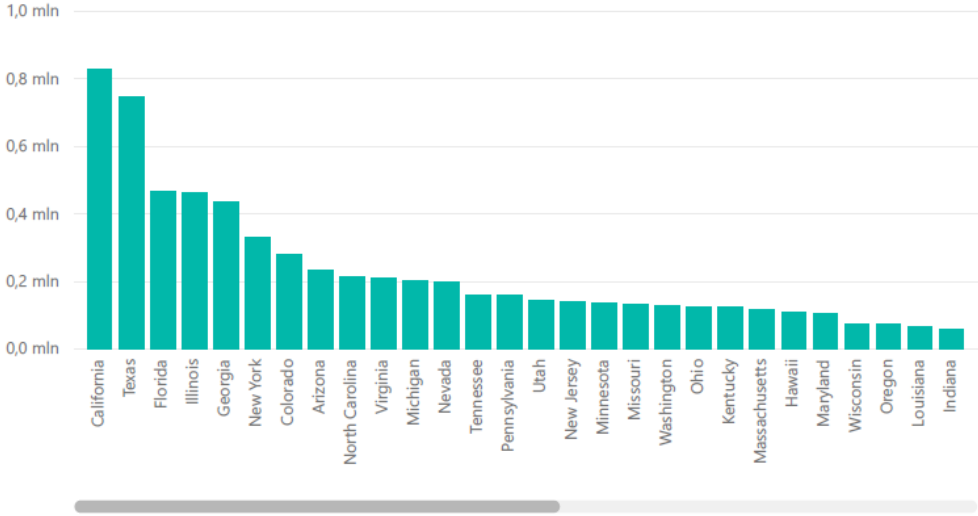
Fact Flight Count wg Airport IATA with more than 50 000 Flights



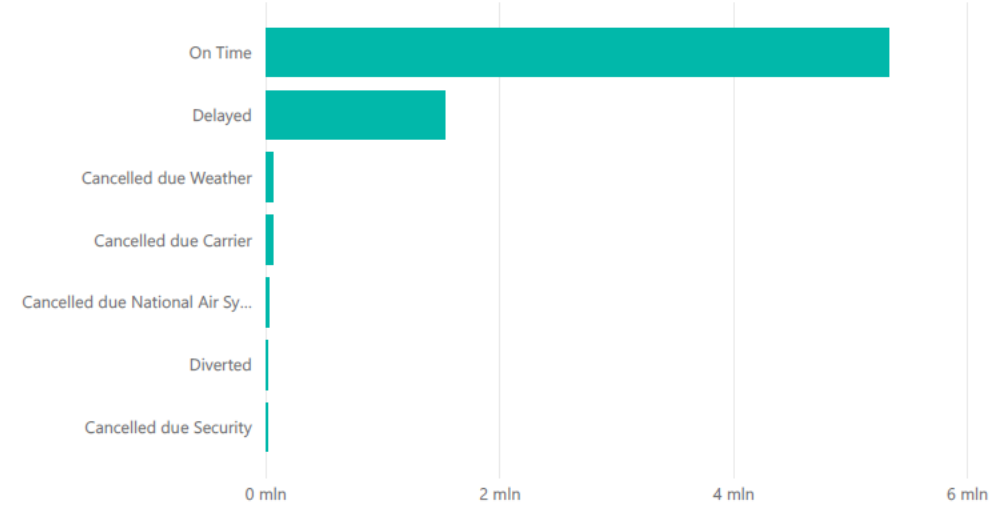
Fact Flight Count wg Airport IATA, Airport Latitude i Airport Longtitude



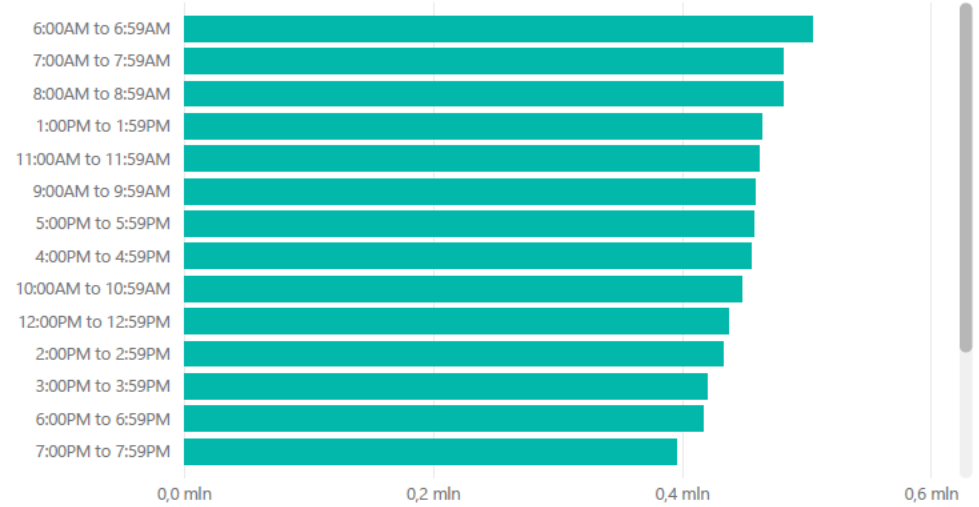
Fact Flight Count wg Airport State Name



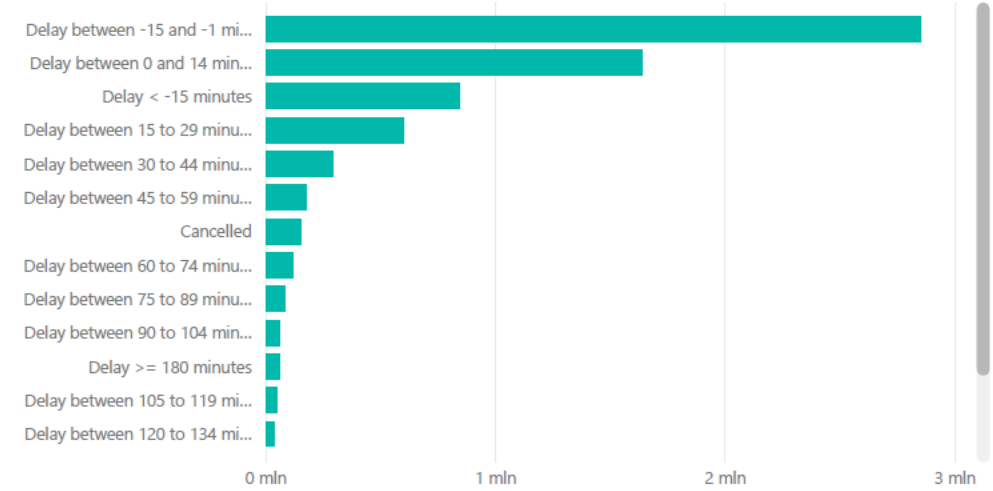
Fact Flight Count wg Flight Status Description



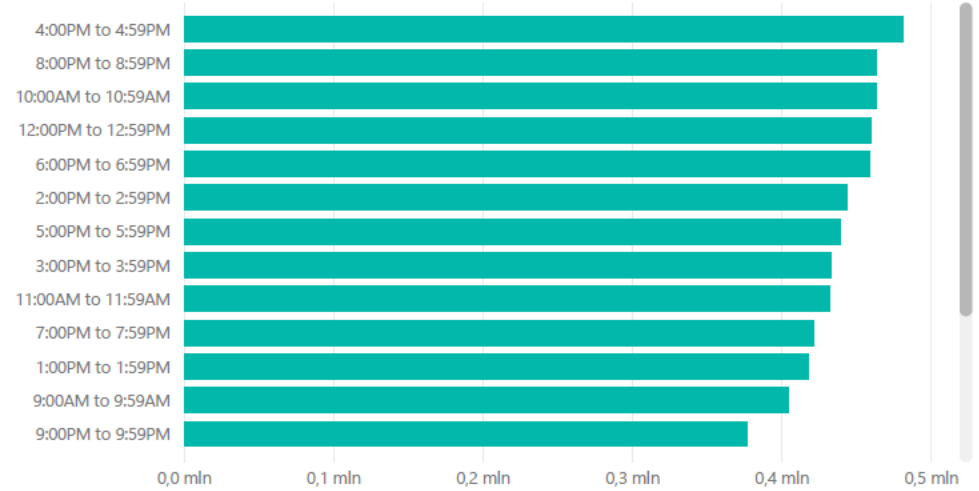
Fact Flight Count wg Departure Time Intervals Description



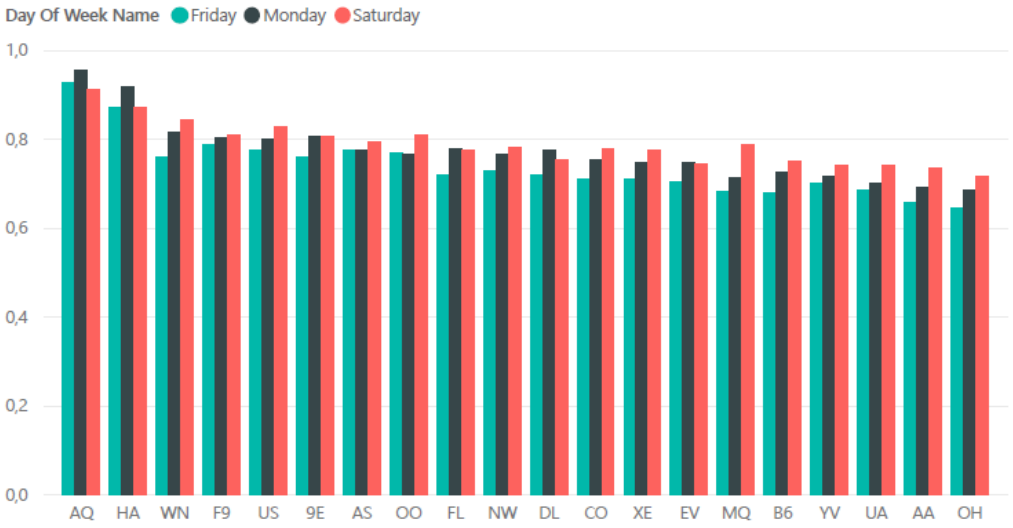
Fact Flight Count wg Arrival Delay Groups Description



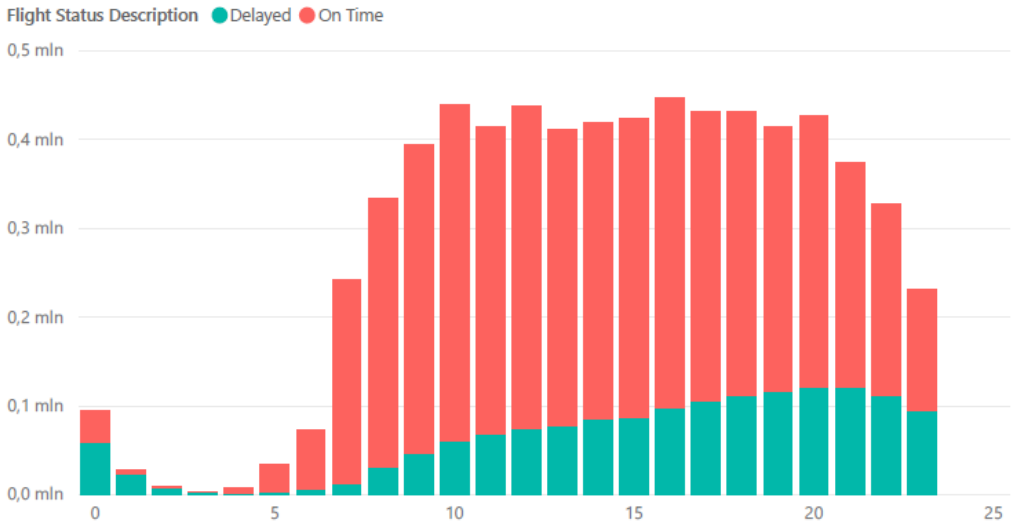
Fact Flight Count wg Arrival Time Intervals Description



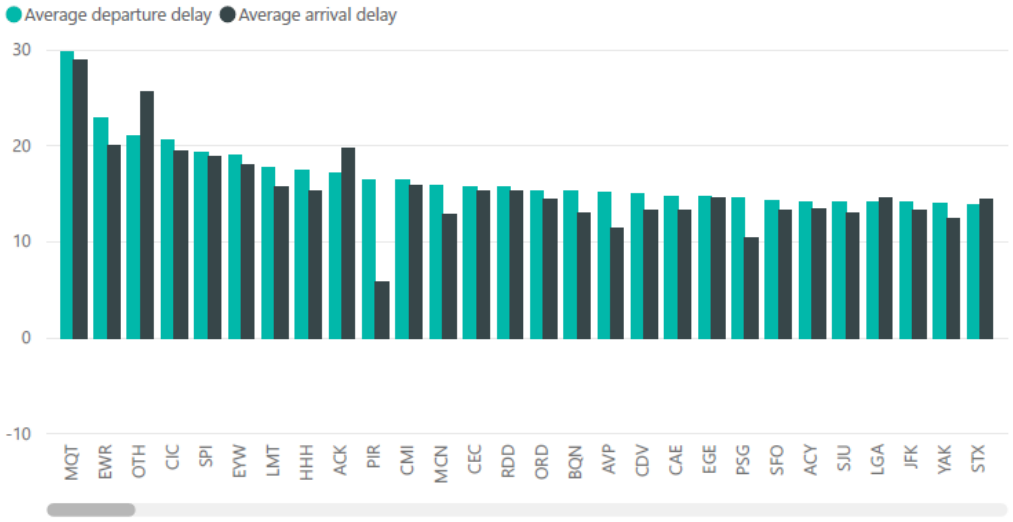
Punctuality wg Airline IATA i Day Of Week Name



Fact Flight Count wg Hour i Flight Status Description



Average departure delay i Average arrival delay wg Airport IATA



Carrier Delay, Nas Delay, Security Delay i Weather Delay