

Wojskowa Akademia Techniczna

Hurtownie Danych

Sprawozdanie z zaliczenia projektowego

Wykonali: Jakub Kapusta, Adam Kochański

Grupa: I6B2S1

Data oddania: 06.06.2019

Prowadzący: dr inż. Marcin Mazurek

1. Analizowane dane

W zadaniu analizowane były dane lotnicze z Ameryki Północnej. Dane zostały pobrane ze stron internetowych:

[Bureau of Transportation Statistics](#)

[Openflights](#)

Za fakt przyjęty został pojedynczy lot.

Tabele wymiarów:

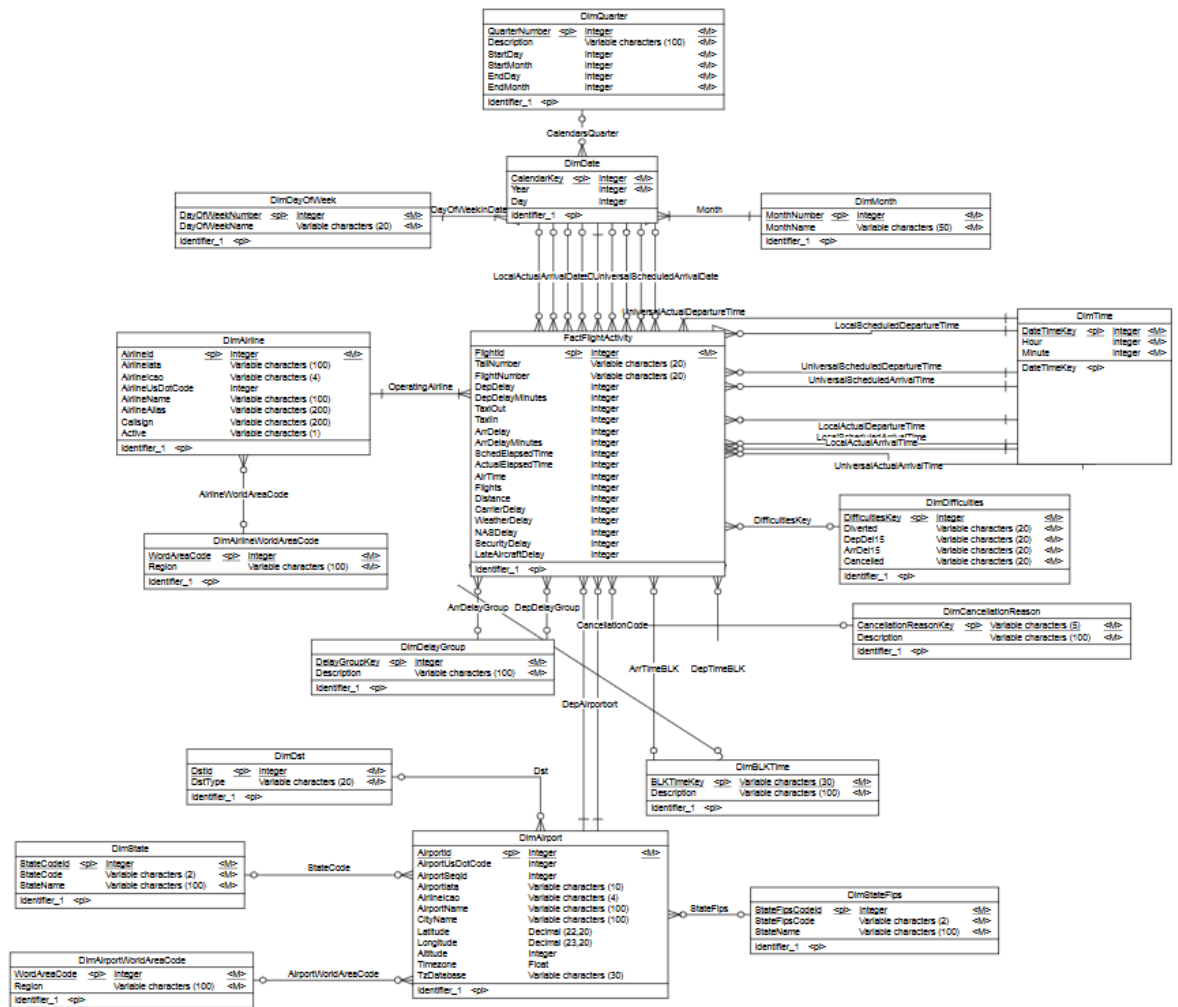
- Date
 - DayOfWeek
 - Month
 - Quarter
- Airline
 - AirlineWorldAreaCode
- Time
- Difficulties
- CancellationReason
- DelayGroup
- BLKTime (Rama czasowa)
- Airport
 - Dst (Region związany z czasem letnim)
 - StateFips (dwucyfrowy kod stanu)
 - StateCode (dwuliterowy kod stanu)
 - AirportWorldAreaCode

Utworzone miary:

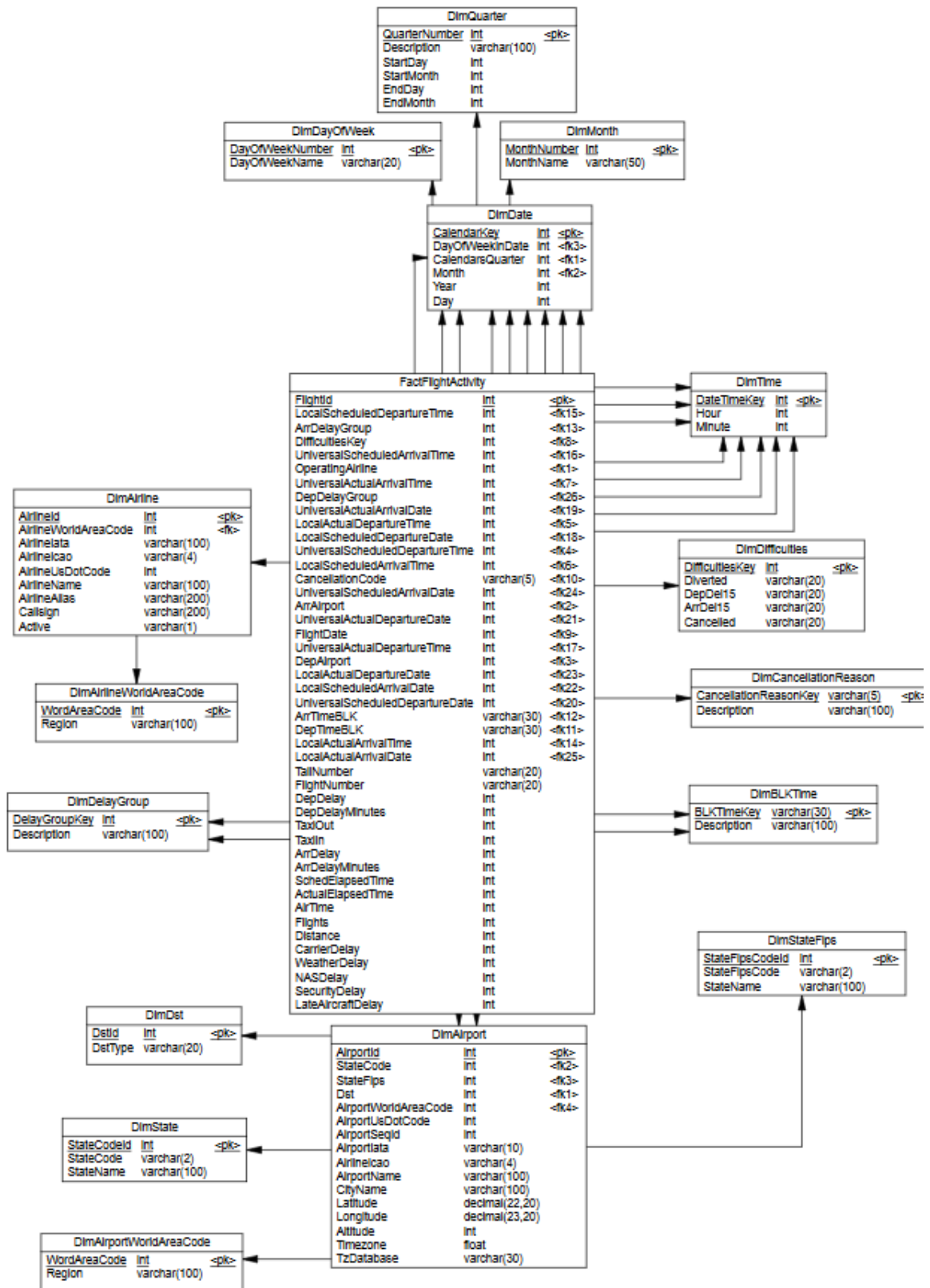
- Przesunięcie czasu wylotu (**DepDelay**) [min] – wartością może być również liczba ujemna oznaczające wcześniejszy start.
- Opóźnienie wylotu (**DepDelayMinutes**) [min] – wartością może być tylko liczba nieujemna.
- Czas kołowania potrzebny do wylotu (**TaxiOut**) [min].
- Czas kołowania potrzebny do przylotu (**TaxiIn**) [min].
- Przesunięcie czasu przylotu (**ArrDelay**) [min] – wartością może być również liczba ujemna oznaczające wcześniejszy przylot.
- Opóźnienie przylotu (**ArrDelayMinutes**) [min] – wartością może być tylko liczba nieujemna.
- Czas spędzony przez samolot w powietrzu (**AirTime**) [min].
- Liczba lotów (**Flights**).
- Dystans (**Distance**) [km].
- Opóźnienie spowodowane przez przewoźnika (**CarrierDelay**) [min].
- Opóźnienie spowodowane przez pogodę (**WeatherDelay**) [min].
- Opóźnienie spowodowane przez National Air System (**NASDelay**) [min].

- Opóźnienie spowodowane przez względy bezpieczeństwa (**SecurityDelay**) [min].
- Opóźnienie spowodowane przez spóźniony samolot (**LateAircraftDelay**) [min].

2. Model konceptualny hurtowni danych



3. Model fizyczny hurtowni danych



4. Skrypty instalacyjne oraz wgrywanie danych

Hd.sql – Skrypt tworzący hurtownię danych SQL Server

Stage.sql-Skrypt tworzący bazę danych Stage SQL Server

Rozwiązanie SSIS_Stage.sln – W kolejnych pakietach wgrywane są dane do kolejnych tabeli bazy danych typu STAGE

Rozwiązanie SSIS_hd.sln- W kolejnych pakietach wgrywane są dane do kolejnych tabeli do właściwej hurtowni danych.

UWAGA- wgrywanie danych dotyczących lotów powinno być uruchomione ostatnie.

Proces ETL

Założenia procesu ETL:

- Zmiana wartości NULL kluczy obcych w tabelach wymiarów na wartość klucza obcego odpowiadająca wartości 'Unknown' lub 'Unknown or Bad'.
- Wydzielenie wszystkich statystyk dotyczących utrudnień występujących podczas lotu do wymiaru DimDifficulties, który zawiera wszystkie możliwe kombinacje powodów utrudnień.
- Utworzenie w wymiarach, których źródłowe pliki .csv zawierały tekstową kolumną jednoznacznie identyfikującą rekord (DimState, DimStateFips) sztucznego klucza głównego będącego typem całkowitym.
- Uzupelnienie tabeli faktów o klucze obce do czasów/dat według następujących założeń:
 - LocalScheduledDepartureTime
 -
- Uzupelnienie tabeli faktów w oparciu o funkcje zwracający odpowiedni klucz obcy do wymiarów (posiadające jako argumentu odpowiednie kolumny ze Stage'owej tabeli faktów).
- Złączenie ze sobą danych z dwóch źródeł (Bureau of Transportation Statistics oraz Openflights) w celu stworzenia wymiarów.

Jak wspomniano w założeniach, proces ETL został przeprowadzony w oparciu o procedury oraz funkcje. Skrypty każdej z procedur oraz funkcji przypisane do Stage'a oraz hurtowni znajdują się odpowiednio w folderach STAGE oraz DWH (w katalogu Skrypty SQL).

Tabela 1 przedstawiająca użyte procedury przypisane do Stage'owej bazy danych używane w procesie ETL

Nazwa procedury	Działanie
deleteAllData	Usunięcie danych ze wszystkich tabel w bazie Stage'owej.
insertNotPresentArea	Dodanie do tabeli WorldAreaCode znajdującej się w Stage'u nieistniejącego jeszcze kraju/regionu ze Stage'owej tabeli Airline oraz Stage'owej tabeli lotniska Airports.
insertNotPresentCarriers	Uzupełnienie w Stage'u tymczasowej tabeli UsdotAndIataMap zawierającej takie atrybuty jak: kod linii lotniczej nadany przez US DOT, kod IATA linii lotniczej oraz nazwę linii lotniczej. Przyjmuje jako argumenty kod IATA linii lotniczej oraz jej nazwę. Procedura jest potrzebna do uzupełnienia w wymiarze hurtowni DimAirline wszystkich informacji o nadanych jej kodach.

Tabela 2 przedstawiająca użyte procedury przypisane do bazy danych hurtowni używane w procesie ETL

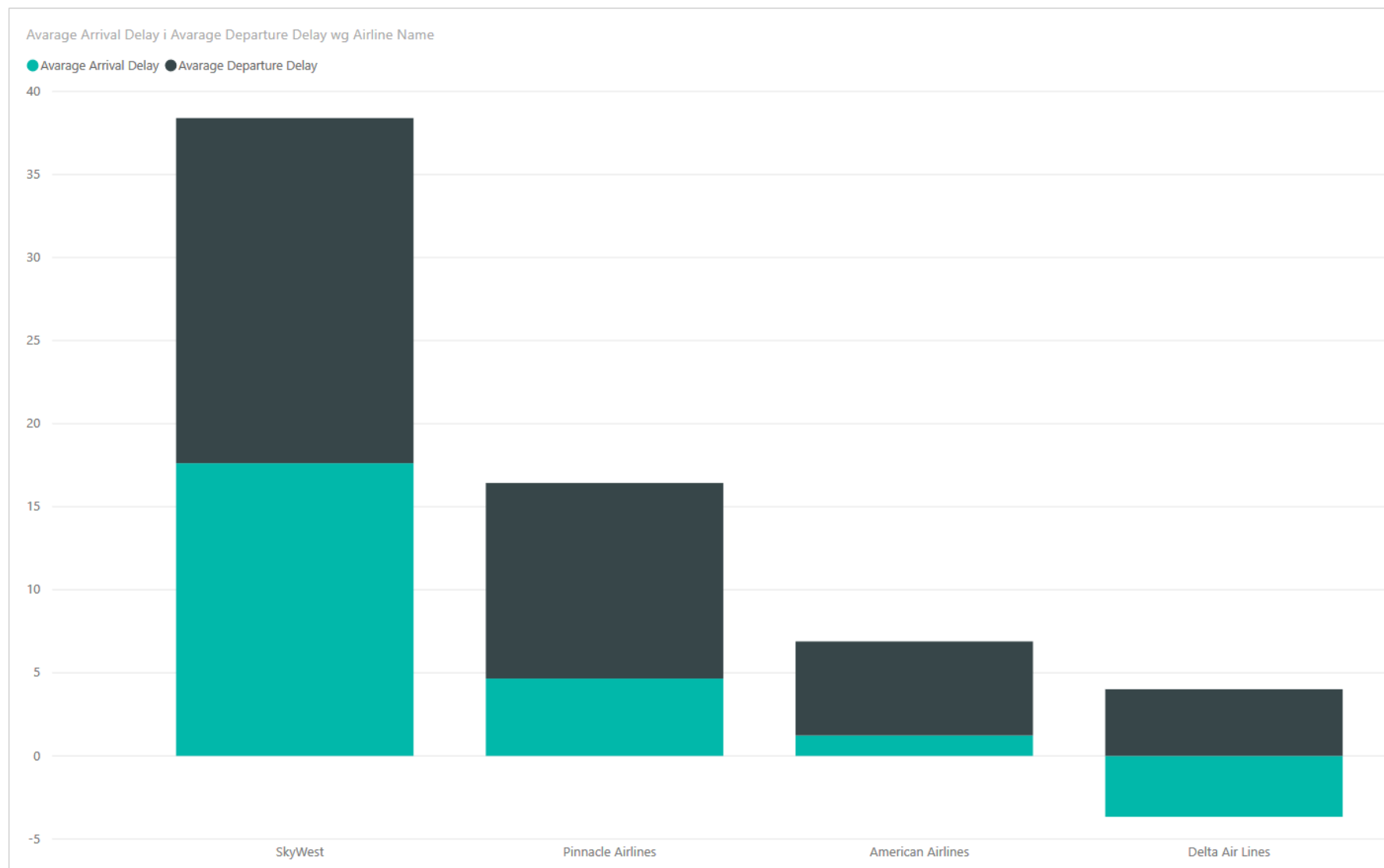
Nazwa procedury	Działanie
getWacCode	Dopasowanie do wymiaru DimAirline odpowiedniego kodu kraju/regionu (WorldAreaCode) na podstawie id linii lotniczej oraz jej kraju siedziby (atrybutów AirlineId oraz Country) Stage'owej tabeli Airlines.
insertNotPresentAirlines	Wstawienie do wymiaru DimAirline nowej linii lotniczej (id, kodu IATA oraz nazwy linii) na podstawie kodu IATA oraz nazwy linii (atrybutów Code oraz Description) Stage'owej tabeli CarrierHistory.
mapUsDotCodeToAirline	Wstawienie do wymiaru DimAirline kodu linii lotniczej nadanego przez US DOT, kodu IATA linii lotniczej oraz jej nazwę (jeśli nie istnieje jeszcze taka linia lotnicza) na podstawie danych z tymczasowej tabeli UsdotAndIataMap znajdującej się w Stage'u. W przypadku, gdy w wymiarze DimAirline znajduje się już linia lotnicza o danym kodzie IATA, następuje zaktualizowanie jej kodu nadanego przez US DOT.
insertNewDstType	Wstawienie do wymiaru DimDst informacji o Dst (Daylight Savings Time), czyli informacji o regionie związanym z czasem letnim lotniska na podstawie jednoliterowego kodu stanowiącego atrybut Dst Stage'owej tabeli Airports.
matchFactWacToAirport	Dopasowanie do wymiaru DimAirport odpowiedniego kodu WorldAreaCode odpowiadającego Stanom Zjednoczonym (United States) na podstawie dostarczonego kodu IATA lotniska ze Stage'owej tabeli faktów. Działanie procedury wynika z tego, że wszystkie fakty pochodzą ze strony Bureau of Transportation Statistics oferującej statystyki lotów odbywających się pomiędzy lotniskami znajdującymi się na terenie Stanów Zjednoczonych.
insertAirportsRows	Wstawienie informacji o lotnisku (atrybuty AirportId , Name , City , Iata , Icao ,Latitude , Longitude , Altitude , TimeZone , Tz ze Stage'owej tabeli Airports) do wymiaru DimAirport. Procedura pomogła rozwiązać konflikty w SSIS związane z rzutowaniem typów zmiennych.

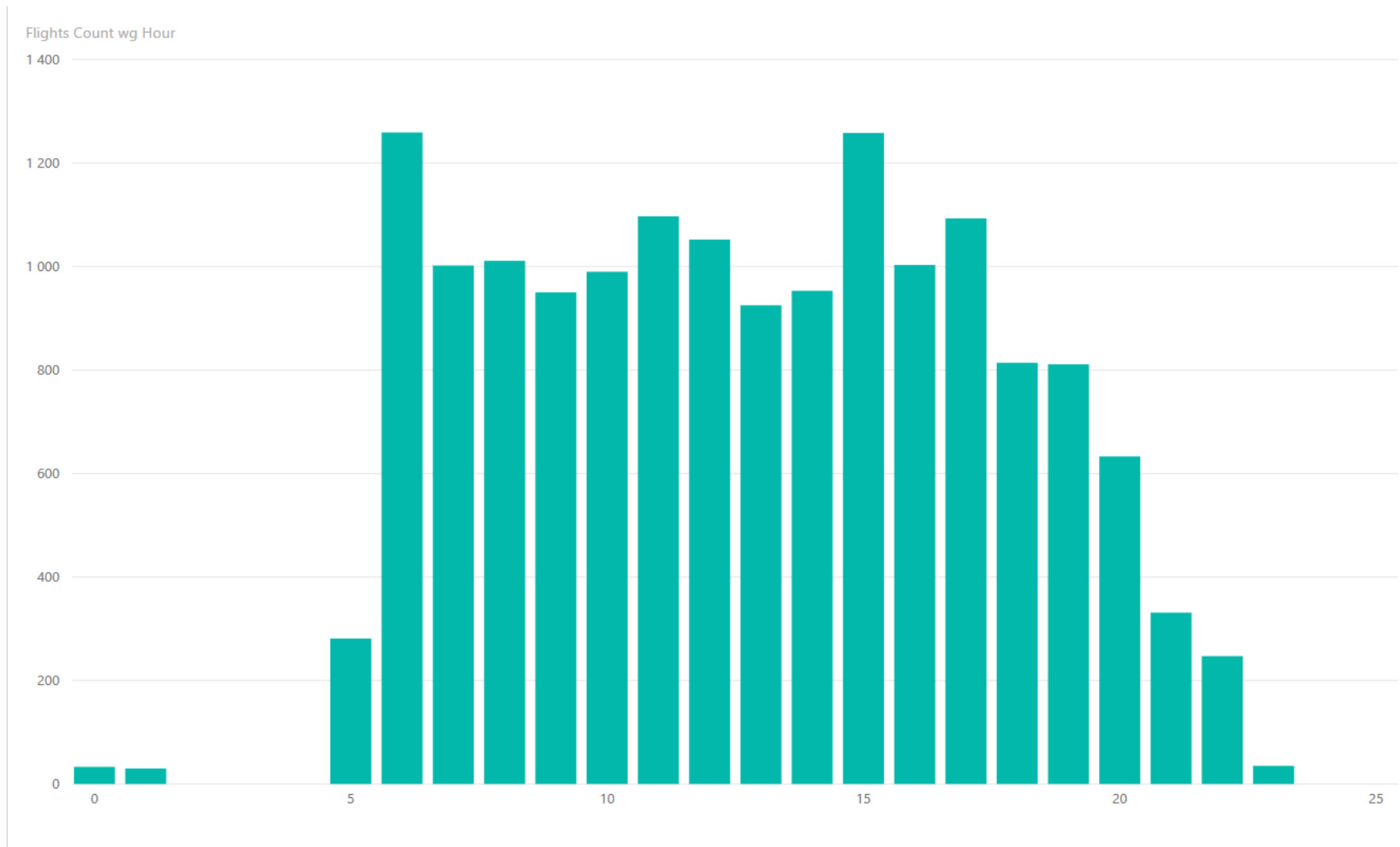
getAirportDstType	Dopasowanie do wymiaru DimAirport odpowiadającego lotnisku regionu Dst na podstawie id lotniska oraz jednoliterowego kodu Dst (atrybutów AirportId oraz Dst) Stage'owej tabeli Airports.
matchAppAirportWacCode	Dopasowanie do wymiaru DimAirport odpowiedniego kodu kraju/regionu (WorldAreaCode) na podstawie id lotniska oraz jej kraju położenia (atrybutów AirportId oraz Country) Stage'owej tabeli Airports.
insertNotPresentAirports	Wstawienie do wymiaru DimAirport nowego lotniska (id, kodu IATA oraz nazwy lotniska) na podstawie kodu IATA oraz nazwy lotniska (atrybutów Code oraz Description) Stage'owej tabeli Airport.
updateAirportCodesAndCities	Zaktualizowanie atrybutów wymiaru DimAirport: kodu nadanego przez US DOT, kodu nadanego przez US DOT (sekwencyjnego) oraz miasta w którym znajduje się lotnisko na podstawie zapytania zwracającego ze Stage'owej tabeli faktów unię zawierającą kod nadany przez US DOT lotnisku startowemu oraz docelowemu, kod nadany przez US DOT (sekwencyjny) lotnisku startowemu oraz docelowemu, kod IATA nadany lotnisku startowemu oraz docelowemu oraz miasta w którym znajduje się lotnisko docelowe oraz startowe.
insertNewState	Wstawienie do wymiaru DimState nieistniejącego jeszcze stanu - jego id, kodu (dwuliterowego) oraz nazwy na podstawie jego kodu oraz nazwy (atrybutów Code, Description) Stage'owej tabeli StateAbrAviation.
insertNewStateFips	Wstawienie do wymiaru DimStateFips nieistniejącego jeszcze stanu - jego id, kodu (dwucyfrowego) oraz nazwy na podstawie jego kodu oraz nazwy (atrybutów Code, Description) Stage'owej tabeli StateFips. Zarówno dla wymiaru DimState oraz DimStateFips zostały wydzielone sztuczne klucze główne w postaci liczby całkowitej)
matchStateCodeToAirport	Przyporządkowanie do wymiaru DimAirport odpowiadającego lotnisku kodu stanu (uaktualnienie kolumn StateCode oraz StateFips) na podstawie zapytania zwracającego ze Stage'owej tabeli faktów unię zawierającą kod nadany przez IATA lotnisku startowemu oraz docelowemu, dwuliterowy kod stanu lotniska startowego oraz docelowego oraz dwucyfrowy kod stanu lotniska startowego oraz docelowego.
insertDates	
insertUnknownDelGr	Dodanie do wymiaru DimDelayGroup nowego rekordu odpowiadającego nieznannej grupie opóźnienia ('Unknown').
insertDifficulties	
insertTime	

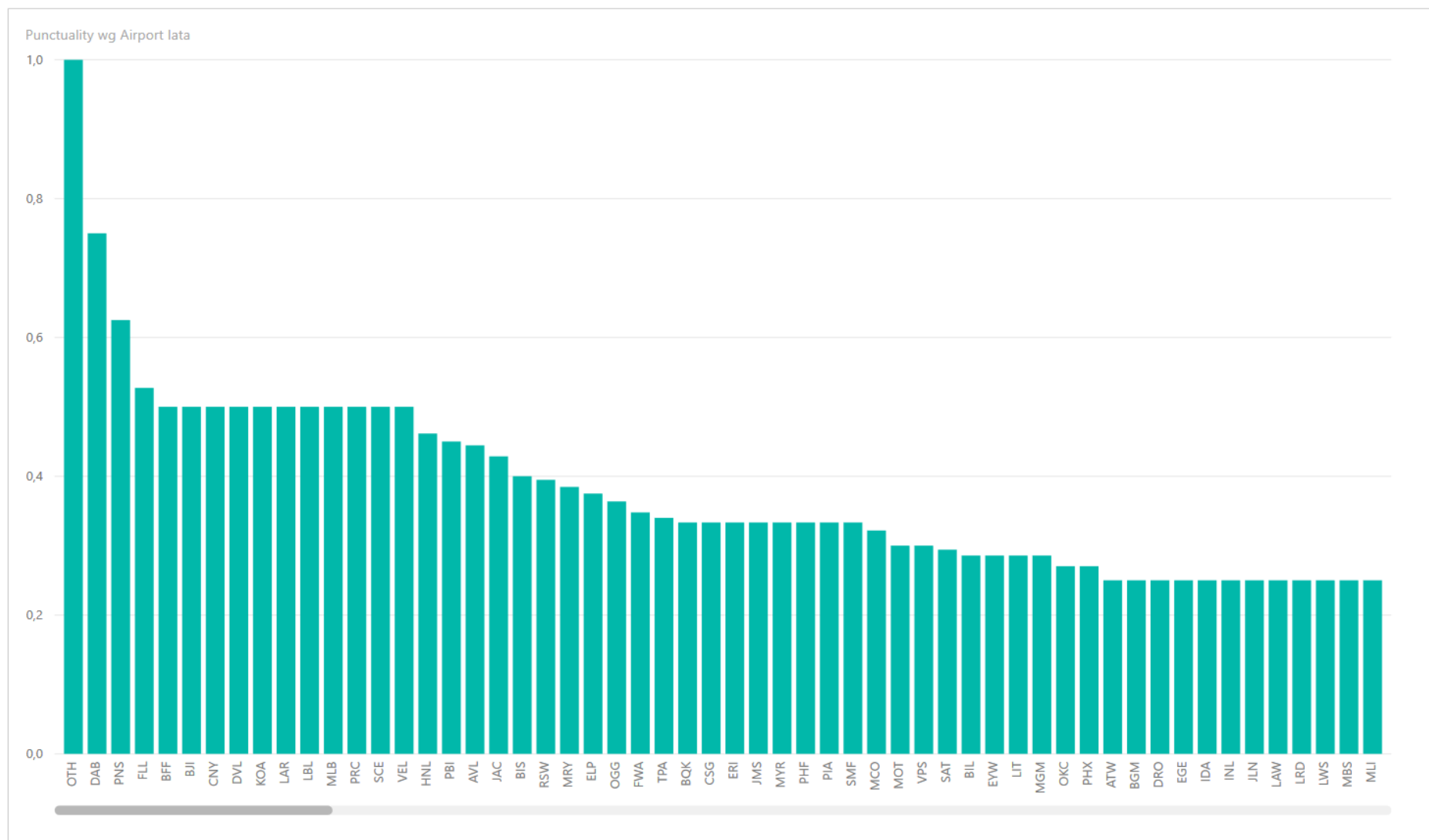
Tabela 3 przedstawiająca użyte funkcje przypisane do bazy danych hurtowni używane w procesie ETL do zasilenia tabeli faktów

Nazwa funkcji	Działanie
getAirlineKey	Dopasowanie do faktu(lotu) odpowiedniego klucza obcego odpowiadającego linii lotniczej (z DimAirline), która realizowała lot (na podstawie kodu IATA linii lotniczej) Stage'owej tabeli faktów.
getAirportKey	Dopasowanie do faktu(lotu) odpowiedniego klucza obcego odpowiadającego lotnisku (z DimAirport), które było lotniskiem startowym/docelowym (na podstawie kodu IATA lotniska startowego/docelowego) Stage'owej tabeli faktów. W związku z tym, że w Stage'owej tabeli faktów istnieje lotnisko startowe oraz docelowe, funkcja w SSIS musiała zostać wywołana dwukrotnie.
getCancellationId	
getDelayGroup	
getDifficultiesKey	
getLocalActualArrivalDate	
getLocalActualDepartureDate	
getLocalScheduledArrivalDate	
getLocalScheduledDepartureDate	
getTimeBLK	
getUniversalActualArrivalDate	
getUniversalActualDepartureDate	
getUniversalScheduledArrivalDate	
getUniversalScheduledDepartureDate	
getUniversalTime	

5. Przykładowe raporty



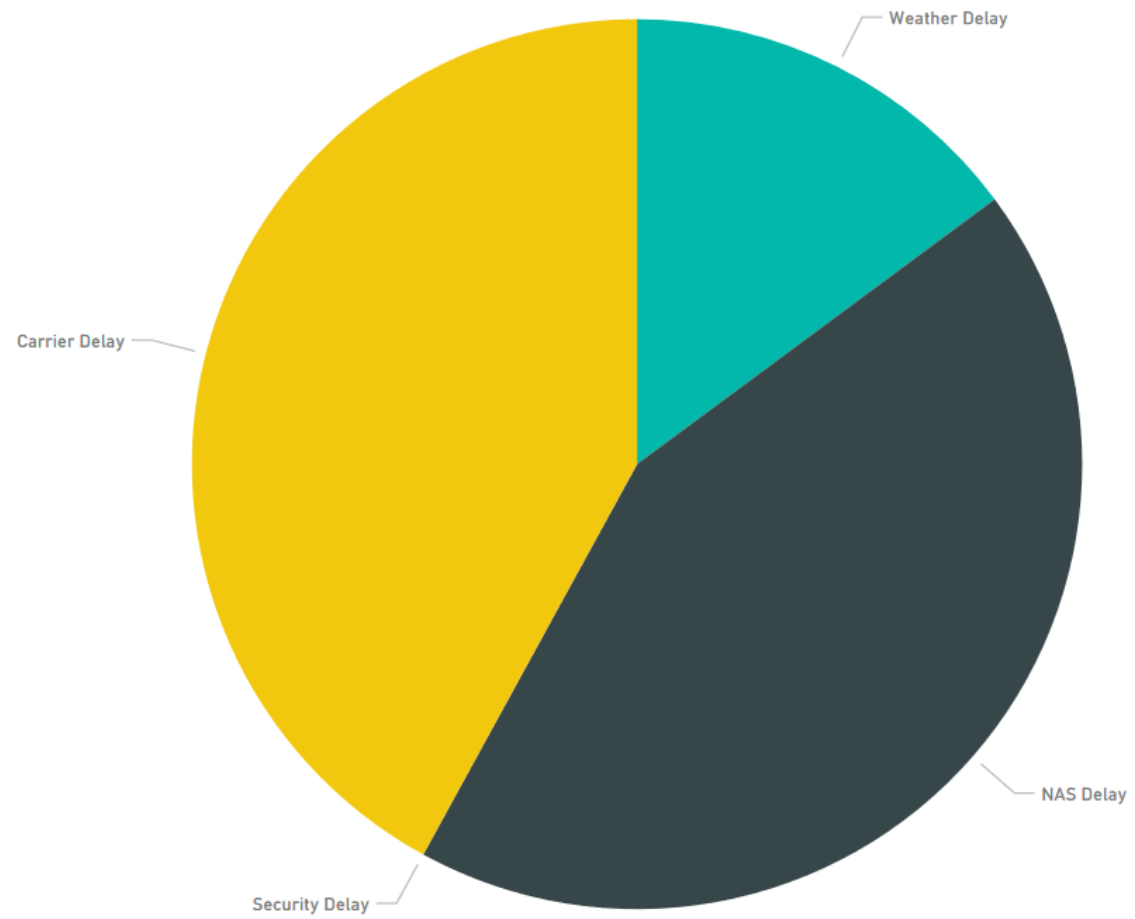


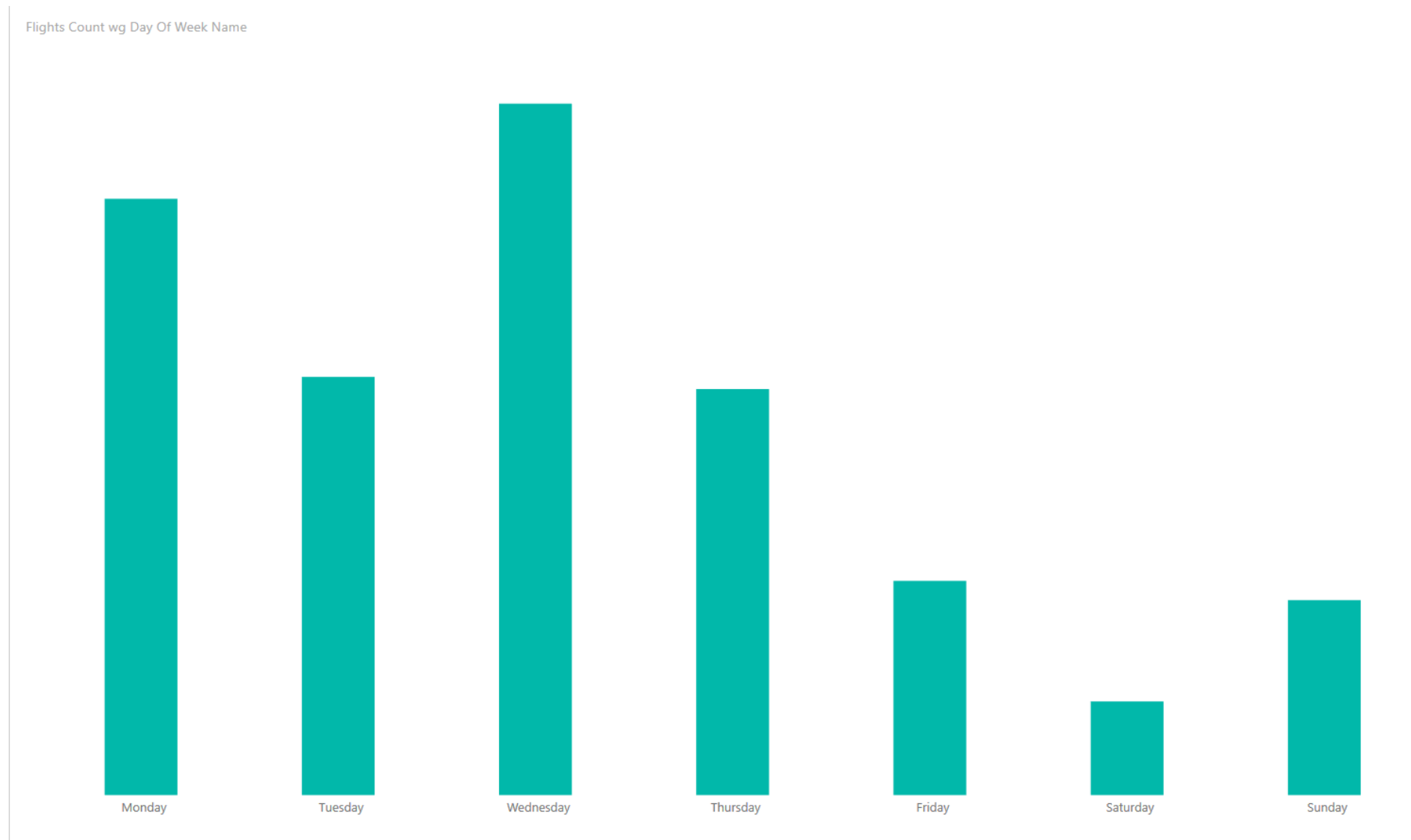


Flights Count wg Airport lata i Airport lata



Weather Delay, NAS Delay, Security Delay i Carrier Delay





Flights Count wg Airport Name

