**Technologie Internetowe**

**Mini Projekt**

**Zestaw danych 9**

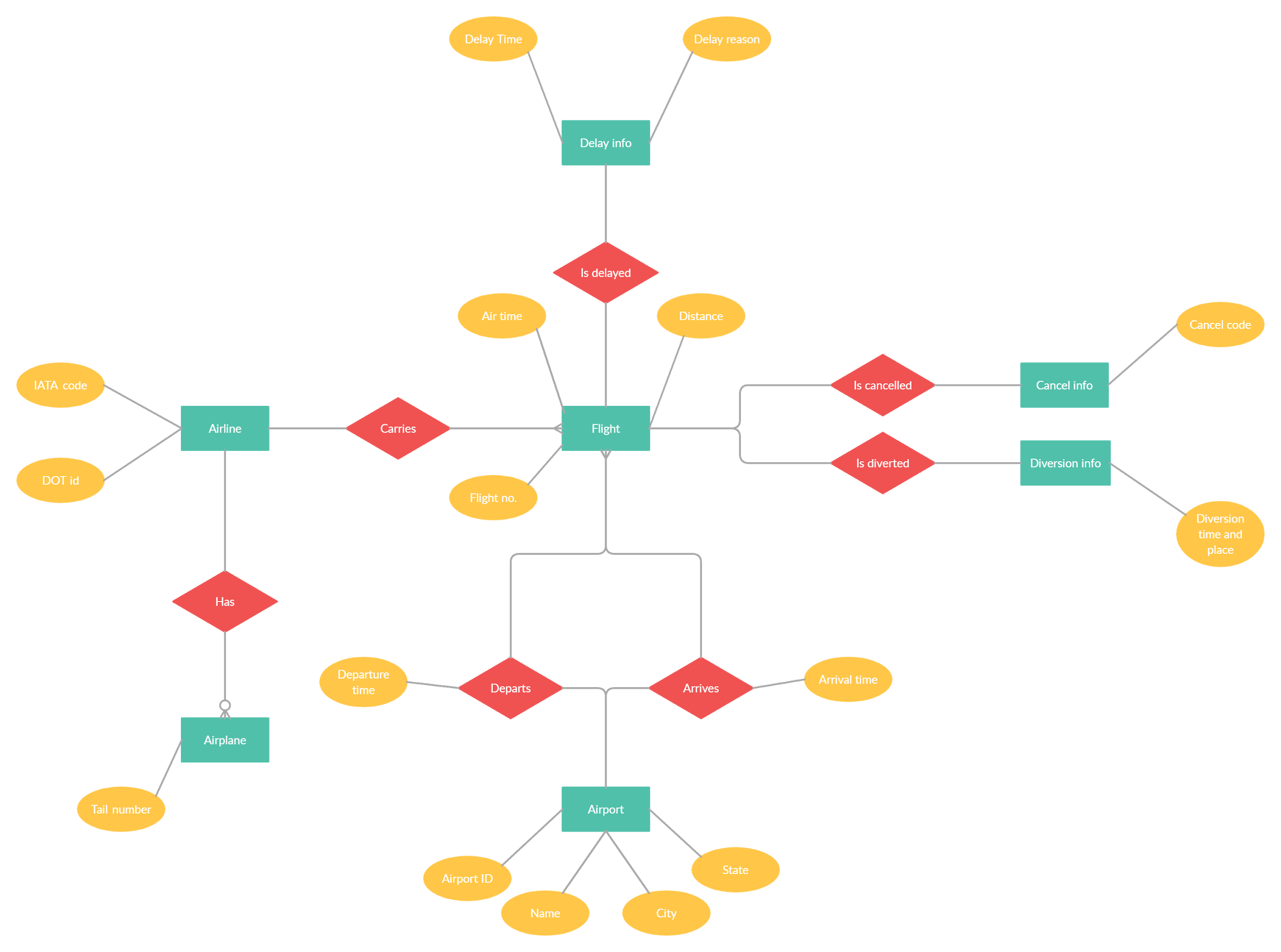
**“On-Time : Reporting Carrier On-Time Performance”**

**Tyszkowski Piotr, Wąsalski Adrian**

***ETAP 1***

1. **Analiza Wymagań**

**Diagram ER**

****

**Użytkownicy systemu:** linie lotnicze, porty lotnicze

**Cel analizy:** Analiza lotów i opóźnień, ich najczęstszych przyczyn i okoliczności w celu ich zapobiegania oraz minimalizacji ich kosztów w przyszłości.

**10 wymagań analitycznych**

Łączny czas opóźnienia na samolot/lotnisko/pilot?.

Średni czas opóźnienia dla linii.

Średni czas opóźnienia dla lotniska.

Średni w powietrzu danej linii.

Ilość obsłużonych połączeń dla lotniska.

Ilość odwołanych lotów na linię.

Najczęstszy powód odwołania lotów.

Powód odwołania zajmujący najwięcej czasu (suma)

Średni czas lotu jednego samolotu

Katastrofy (samolot wyleciał ale nie wylądował)

*Dopisać avg consump do samolotu*

**Proces do analizy:** Lot samolotu.

**Miary:** Efektywność mierzona według czasu przelotu/opóźnienia/kosztów

**Perspektywy analizy:** Rok, miesiąc, dzień, samolot, linia lotnicza, lotnisko, przyczyny opóźnień, przyczyny odwołań, przyczyny przekierowań

**2. Opracowanie modelu wielowymiarowego**

**2.1 Zdefiniowanie faktu**

Za fakt przyjmujemy przelot samolotu, rozumiany jako przelot z lotniska A do lotniska B. Samolot wyleciał i wylądował. Typ faktu określamy jako transakcyjny (transactional) gdyż każde informacje o przelotach są od siebie różne i się nie powtarzają.

Miary oraz sposób ich mierzenia :

Długość lotu w minutach (Airtime) - nieaddytywna

Odległość lotu w kilometrach (Distance) -nieaddytywna

Opóźnienie lotu w minutach (Couse\_of\_delay) - nieaddytywna

Ilość wykonanych lotów (suma z Flight summaries) - addytywna

**2.2 Wybór wymiarów**

Cause\_of\_delay - informacje o przyczynie opóżnienia lotu oraz o długości tego opóźnienia w minutach

Airline - Linia lotnicza obsługująca lot

Origin airport - lotnisko z którego wyleciał samolot

Destination airport - lotnisko do którego leciał samolot

Flight\_date - data lotu

|  |
| --- |
| Cancelations |
| cad\_id |
| [Cancelle](https://www.transtats.bts.gov/FieldInfo.asp?Field_Desc=Cancelled%20Flight%20Indicator%20%281%3DYes%29&Field_Type=Num&Lookup_Table=L_YESNO_RESP&Table_ID=236&SYS_Table_Name=T_ONTIME_REPORTING&Sys_Field_Name=CANCELLED)ationReason |
| [CancellationCode](https://www.transtats.bts.gov/FieldInfo.asp?Field_Desc=Specifies%20The%20Reason%20For%20Cancellation&Field_Type=Char&Lookup_Table=L_CANCELLATION&Table_ID=236&SYS_Table_Name=T_ONTIME_REPORTING&Sys_Field_Name=CANCELLATION_CODE) |

|  |
| --- |
| Flight\_date |
| Date |
| year |
| [Month](https://www.transtats.bts.gov/FieldInfo.asp?Field_Desc=Month&Field_Type=Num&Lookup_Table=L_MONTHS&Table_ID=236&SYS_Table_Name=T_ONTIME_REPORTING&Sys_Field_Name=MONTH) |
| DayofMonth |
| [DayOfWeek](https://www.transtats.bts.gov/FieldInfo.asp?Field_Desc=Day%20of%20Week&Field_Type=Num&Lookup_Table=L_WEEKDAYS&Table_ID=236&SYS_Table_Name=T_ONTIME_REPORTING&Sys_Field_Name=DAY_OF_WEEK) |

|  |
| --- |
| Airline |
| Airline id |
| Air\_carrier\_code |
| name |

|  |
| --- |
| DelayInfo |
| Delay\_id |
| CarrierDelay |
| WeatherDelay |
| NASDelay |
| SecurityDelay |
| LateAircraftDelay |

|  |
| --- |
| (Origin/Destination) Airport |
| AirportID |
| Airport\_state |
| Airport |

**2.3** **Poziom szczegółowości danych**

1. Dla tabeli faktu

Agregowanie na poziomie lotu, jeden fakt reprezentuje jeden skończony lot

1. Dla wymiarów

**Cause\_of\_delay** - jeden byt opisuje powód opóźnienia (jeżeli zaszło) oraz czas opóźnienia dla danego lotu

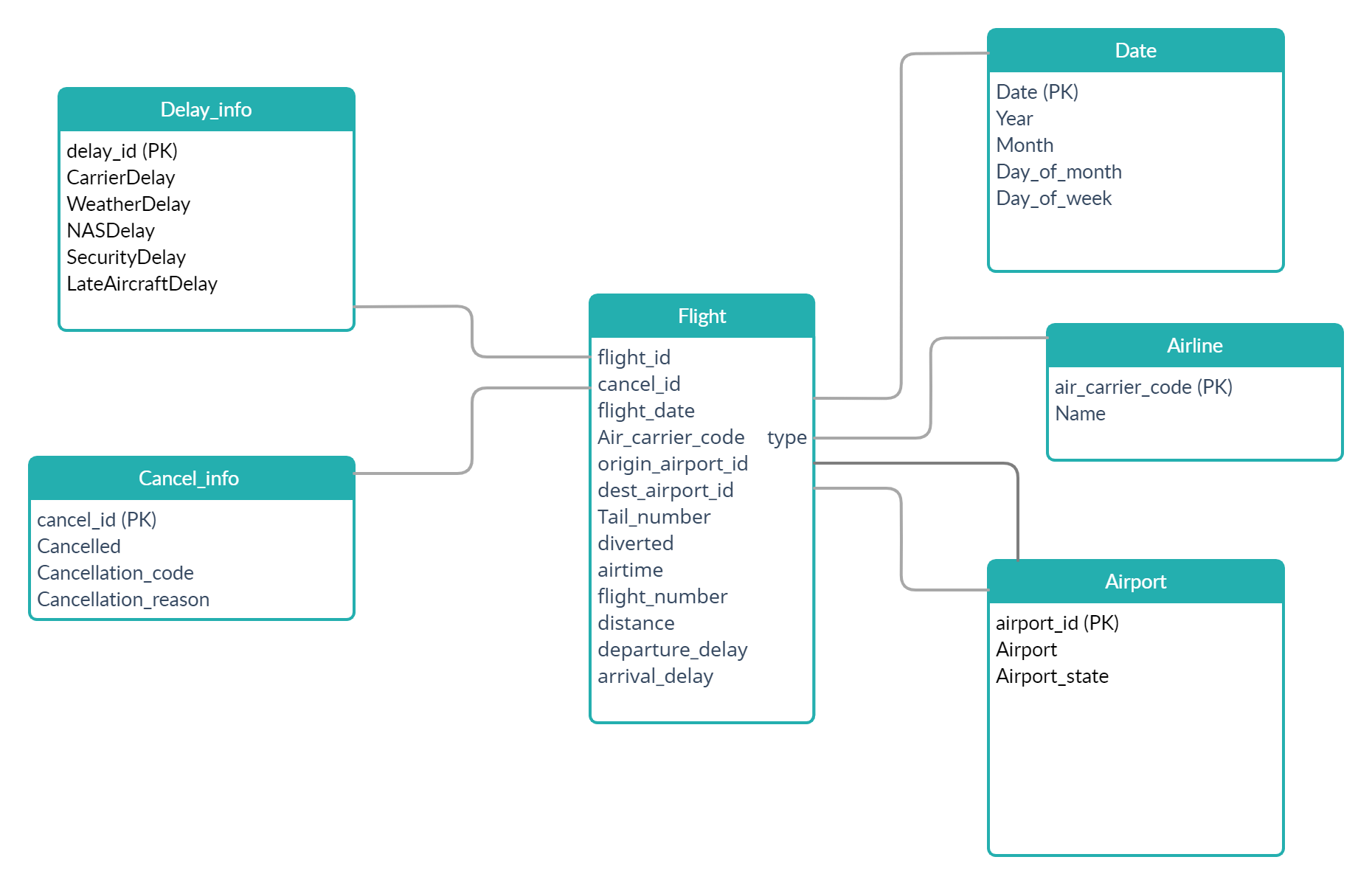
**Airline** - jeden byt opisuje konkretną linię

**Airport** - jeden byt opisując jedno konkretne lotnisko

**Delay info** - jeden byt opisujący powody opóźnienia

**Cancel info** - jeden byt opisuje odwołanie ( jeżeli zaszło)

**2.4 Model wielowymiarowy danych (model gwiazdy)**



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model docelowy** | | | **Dane źródłowe** | | | **Transformacja** |
| **Tabela** | **Kolumna** | **Typ** | **Tabela** | **Kolumna** | **Typ** |  |
| Flights | FlightID | int | - | - | - | To be added |
| Flights | Flight number | int | raw\_data | FL\_NUM | int |  |
| Flights | FlightDate | date | raw\_data | FL\_DATE | date |  |
| Flights | AirCarrierCode (airlineid) | char(2) | raw\_data | UNIQUE\_CARRIER | char(2) |  |
| Flights | OriginAirportID | int | raw\_data | ORIGIN\_AIRPORT\_ID | int |  |
| Flights | DestinationAirportID | int | raw\_data | DESTINATION\_AIRPORT\_ID | int |  |
| Flights | TailNumber | varchar | raw\_data | TAIL\_NUM | string |  |
| Flights | AirTime | int | raw\_data | AIR\_TIME | int |  |
| Flights | Distance | int | raw\_data | DISTANCE | int |  |
| Flights | Distance group | int | raw\_data | DISTANCE\_GROUP | int |  |
| Flights | Canceled | bool | raw\_data | CANCELLED | int | 1 = canceled, 0 = not cancelled |
| Flights | CancelD | int | - | - | - | If cancelled: add cancel id  Else: NULL |
| Flights | Diverted | bool | raw\_data | DIVERTED | int | 1 = diverted, 0 = not diverted |
| Flights | DepartureDelay | int | raw\_data | DEP\_DELAY | int |  |
| Flights | ArrivalDelay | int | raw\_data | ARR\_DELAY | int |  |
|  |  |  |  |  |  |  |
| Dates | Date | int | raw\_data | FL\_DATE | date | Unique FL\_DATE |
| Dates | Year | int | raw\_data | YEAR | int |  |
| Dates | Month | int | raw\_data | MONTH | int |  |
| Dates | Day of month | int | raw\_data | DAY\_OF\_MONTH | int |  |
| Dates | Day of week | varchar | raw\_data | DAY\_OF\_WEEK | int | 1,2,3,4,5,6,7 = pn,wt,śr... |
|  |  |  |  |  |  |  |
| Airports | AirportID | int | raw\_data | ORIGIN\_AIRPORT\_ID, DEST\_AIRPORT ID |  | Unique ORIGIN\_AIRPORT\_ID + DEST\_AIRPORT ID |
| Airports | Airport | char(3) | raw\_data | ORIGIN, DEST |  | Unique + ORIGIN, DESTINATION |
| Airports | AirportState | char(3) | raw\_data | ORIGIN\_STATE\_ABR, DEST\_STATE\_ABR |  |  |
|  |  |  |  |  |  |  |
| Cancel Info | CancelID | int | - | - | - | Flights FK |
| Cancel Info | CancellationCode | int | raw\_data | CANCELLATION\_CODE | char | A,B,C,D |
| Cancel Info | CancellationReason | varcahr | raw\_data | - | - | A = carrier, B = weather, C = NAS, D = security |
|  |  |  |  |  |  |  |
| DelayInfo | FlightID | int | - | - | - | Flight FK |
| DelayInfo | CarrierDelay | int | raw\_data | CARRIER\_DELAY |  |  |
| DelayInfo | WeatherDelay | int | raw\_data | WEATHER\_DELAY |  |  |
| DelayInfo | NASDelay | int | raw\_data | NAS\_DELAY |  |  |
| DelayInfo | SecurityDelay | int | raw\_data | SECURITY\_DELAY |  |  |
| DelayInfo | LateAircraftDelay | int | raw\_data | LATE\_AIRCRAFT\_DELAY |  |  |
|  |  |  |  |  |  |  |
| Airlines | AirCarrierCode | char | raw\_data | UNIQUE\_CARRIER | char |  |
| Airlines | CarrierName | varchar |  | - | - | https://www.census.gov/foreign-trade/reference/codes/aircarrier/acname.txt |

<https://www.census.gov/foreign-trade/reference/codes/aircarrier/acname.txt>

<https://fred.stlouisfed.org/series/DCOILBRENTEU>

***ETAP 2***

1. **Projekt procesu ETL**

1.1 **Projekt procesu ekstrakcji oraz transformacji danych**

1. Sposób dostępu do danych źródłowych

* Główne źródło danych

Dostęp do danych został uzyskany dzięki stronie internetowej Biura statystyki transportowej (Bureau of transportation statistics). Na tej stronie możliwe było pobranie danych na temat odbytych lotów od roku 1988 do roku 2012. Dane są dostępne w plikach csv.

https://www.transtats.bts.gov/nosessionvar.asp

* Dane uzupełniające

Ze względu na brakujące dane konieczne było skorzystanie z następujących źródeł danych uzupełniających:

Dane o cenach ropy - pobrane zostały ze , w formacie csv.

https://fred.stlouisfed.org/series/DCOILBRENTEU?fbclid=IwAR1sX1vewxPndk1EieMr9UcC\_KUcK2kPD3l52VM5w9ayusOfVRsUB38M1bM

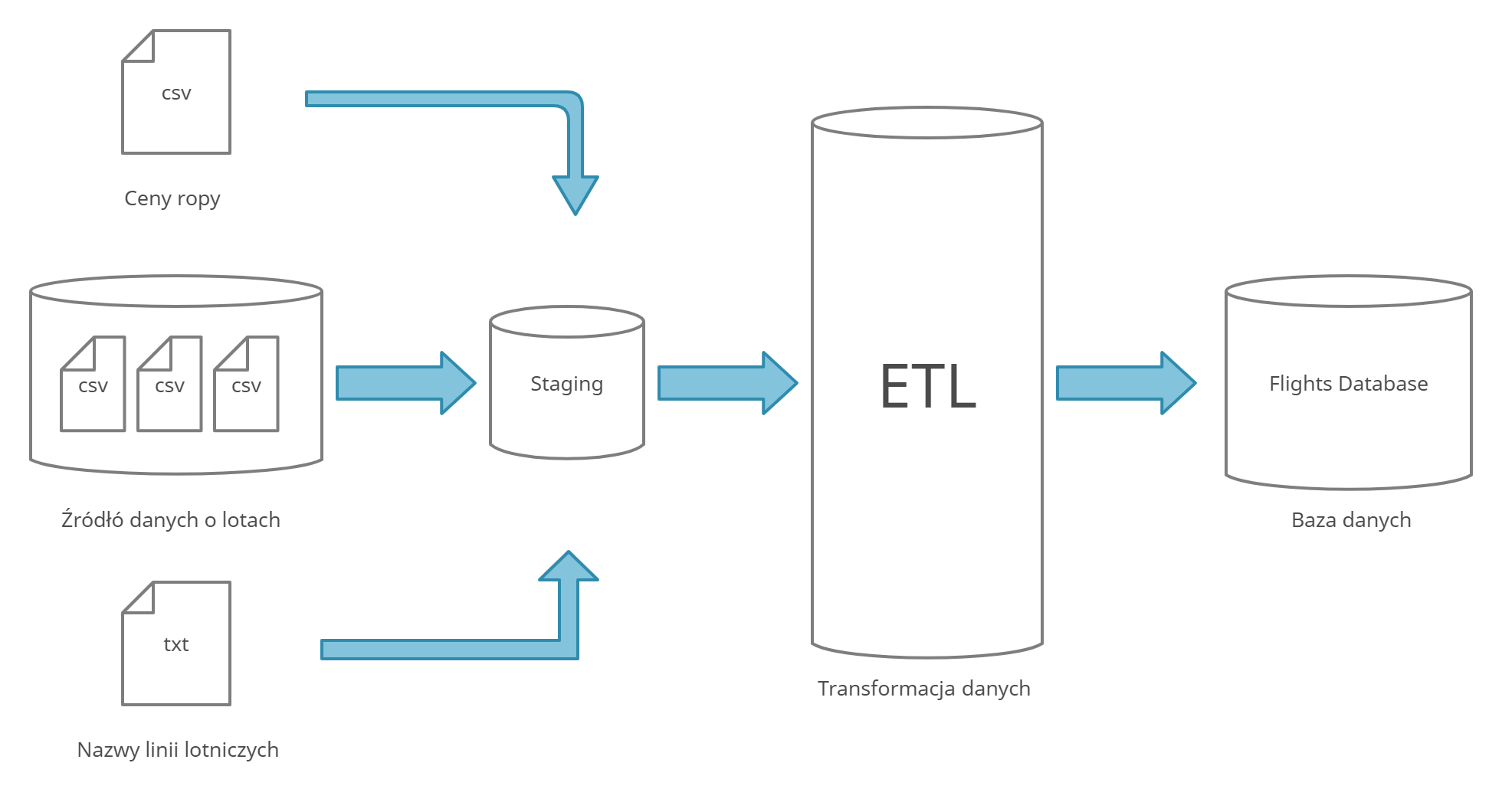
Dane o rozwinięciu akronimów linii lotniczych - konieczne w przejrzystym analizowaniu wyników jako plik txt.

<https://www.census.gov/foreign-trade/reference/codes/aircarrier/acname.txt?fbclid=IwAR0gx1UfU9q1pEAt5Pq5XpcyPdYbD-g-gKdwnNaunRWtpku8AoeCRXXKZ0U>

1. Potrzebne transformacje

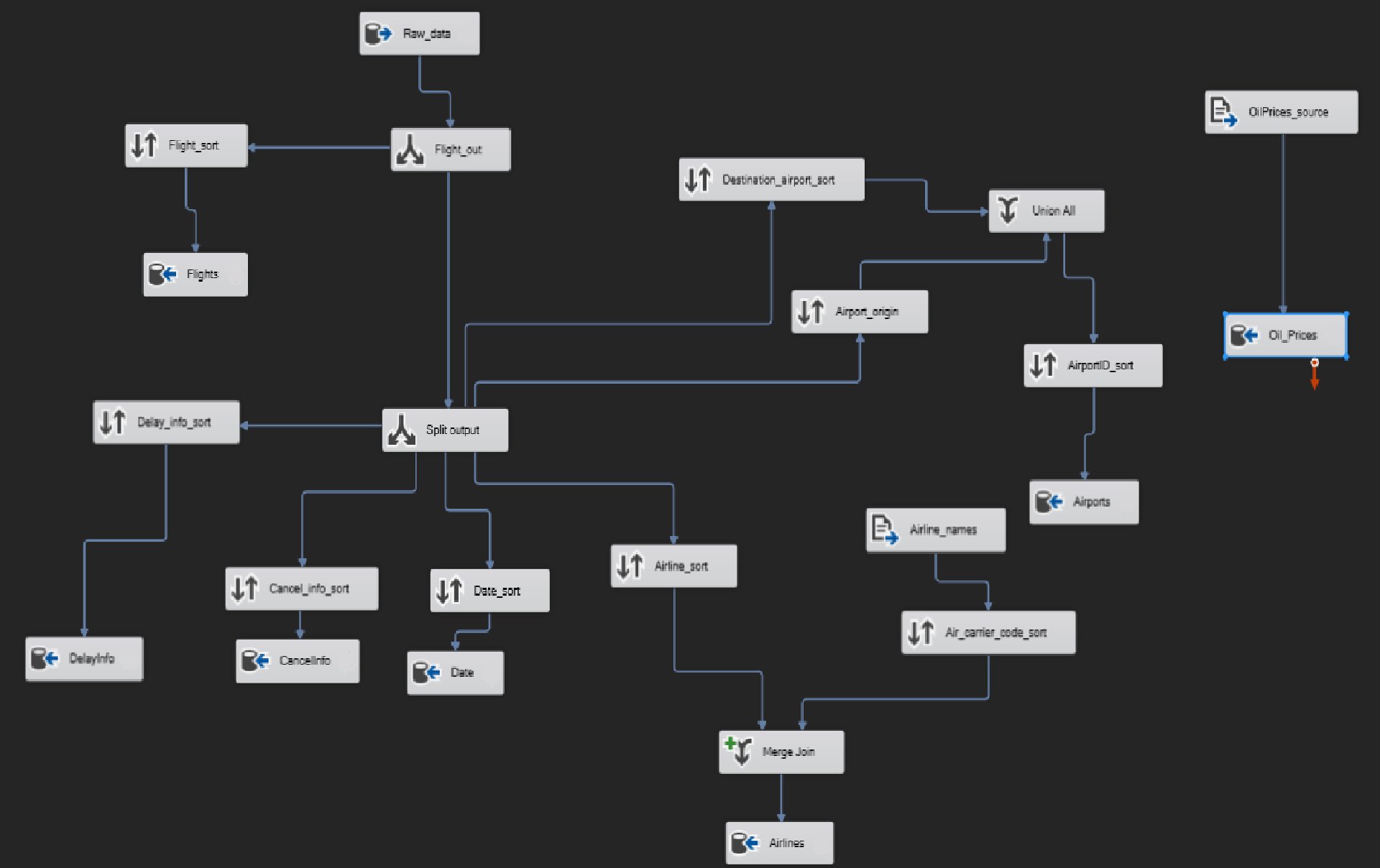
Dane źródłowe są dostępne jako jedna wielka tabela wszystkich rekordów o jednym typie danych (Varchar). Z tego powodu konieczne jest sortowanie kolumn oraz określenie odpowiednich typów danych zgodnie z utworzoną wcześniej mapą logiczną. Sortowanie w tym przypadku polega na przydzieleniu odpowiednich danych do odpowiednich kolumn tak jak przedstawia nasz model wielowymiarowy. W trakcie tego podziału również uzupełniamy bazę o brakujące dane z zewnętrznych źródeł. Konieczne również było uzupełnienie danych o klucze główne i obce (Flight\_id). Ze względów technicznych skorzystaliśmy tylko z ostatniego roku dostępnego w danych źródłowych (2012), dzięki temu udało się zmniejszyć ilość rekordów z 70 mln do około 6 mln.

1.2 **Prosty schemat procesu ETL**



2. **Implementacja procesu jako pakiet SSIS**

Prezentacja przygotowanego pakietu SSIS w którym dokonuje się ekstrakcja danych, podstawowe czyszczenie danych oraz przygotowanie danych do analizy. Poniżej dostępny jest podgląd zakładki Data Flow w tym pakiecie. Pakiet SSIS dostępny jest w załączniku.



***ETAP 3***

1. Weryfikacja poprawności danych

Przeprowadzając weryfikację statystyk wymiarów potwierdziliśmy, że liczba rekordów oraz zakres danych pomiędzy danymi źródłowymi a po przekształceniach zgadzają się ze sobą.

1. Miary obliczane

2.1 Miary obliczane

Łączny wskaźnik kosztów lotu - (distance \* oil\_price)

Łączny czas opóźnień

2.2 KPI

* Średni czas opóźnienia - Informuje nas jak bardzo spóźnia się dana linia, dane lotnisko itp.
* Procent opóźnionych lotów (o więcej niż 15 minut) - Informuje nas ile lotów jest opóźnionych

1. Dostęp i agregacje

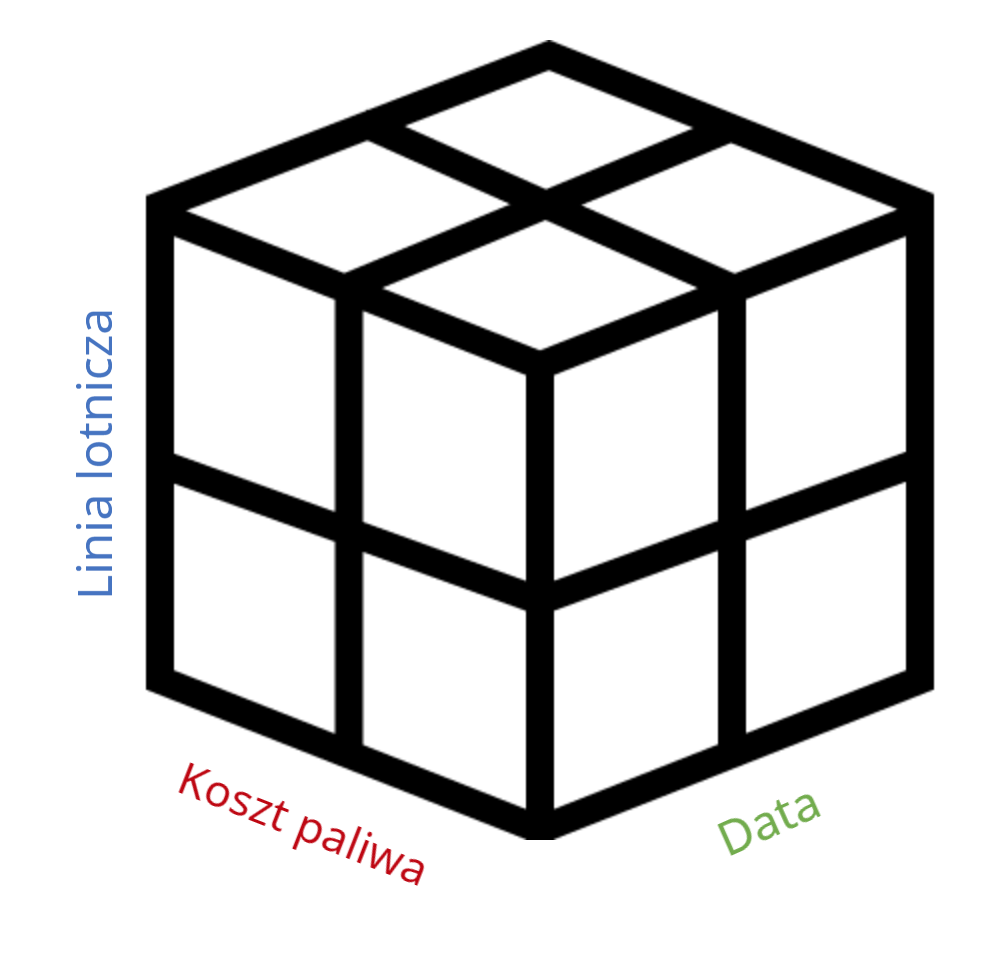
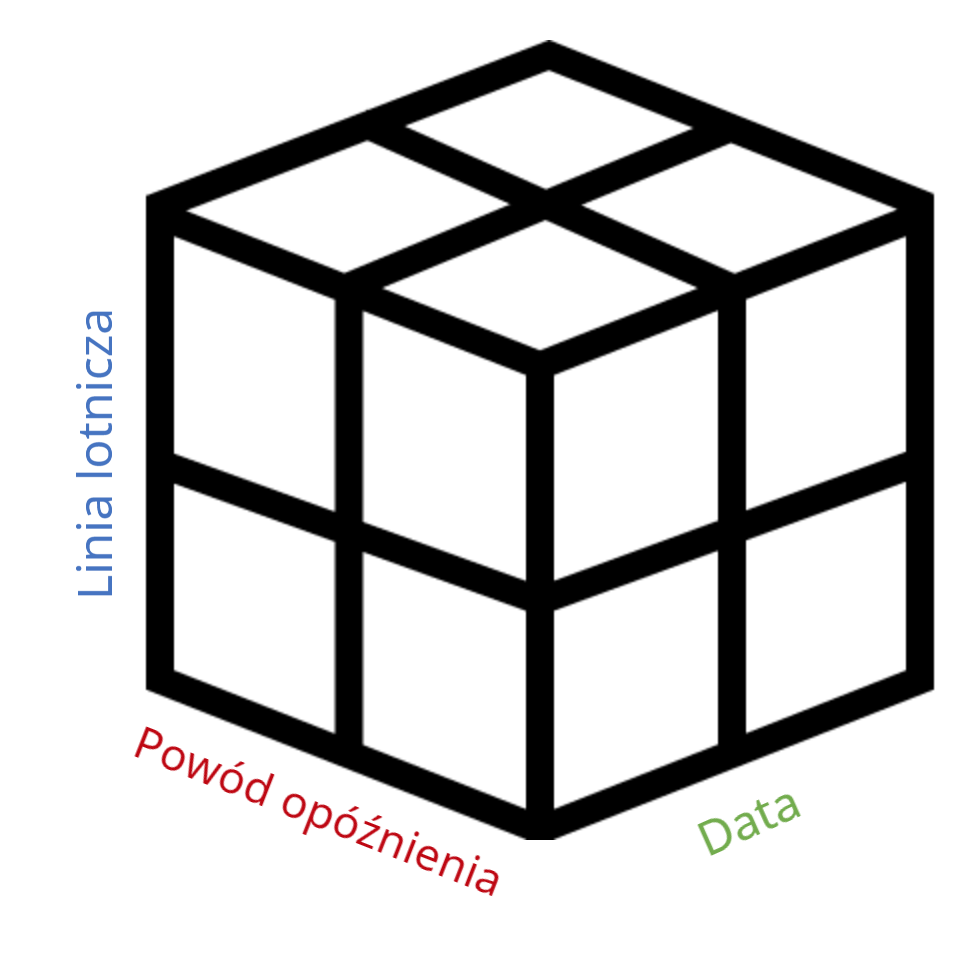
3.1 Perspektywy dostępu do danych

* Lotnisko
* Linia lotnicza
* Miesiąc
* Kwartał

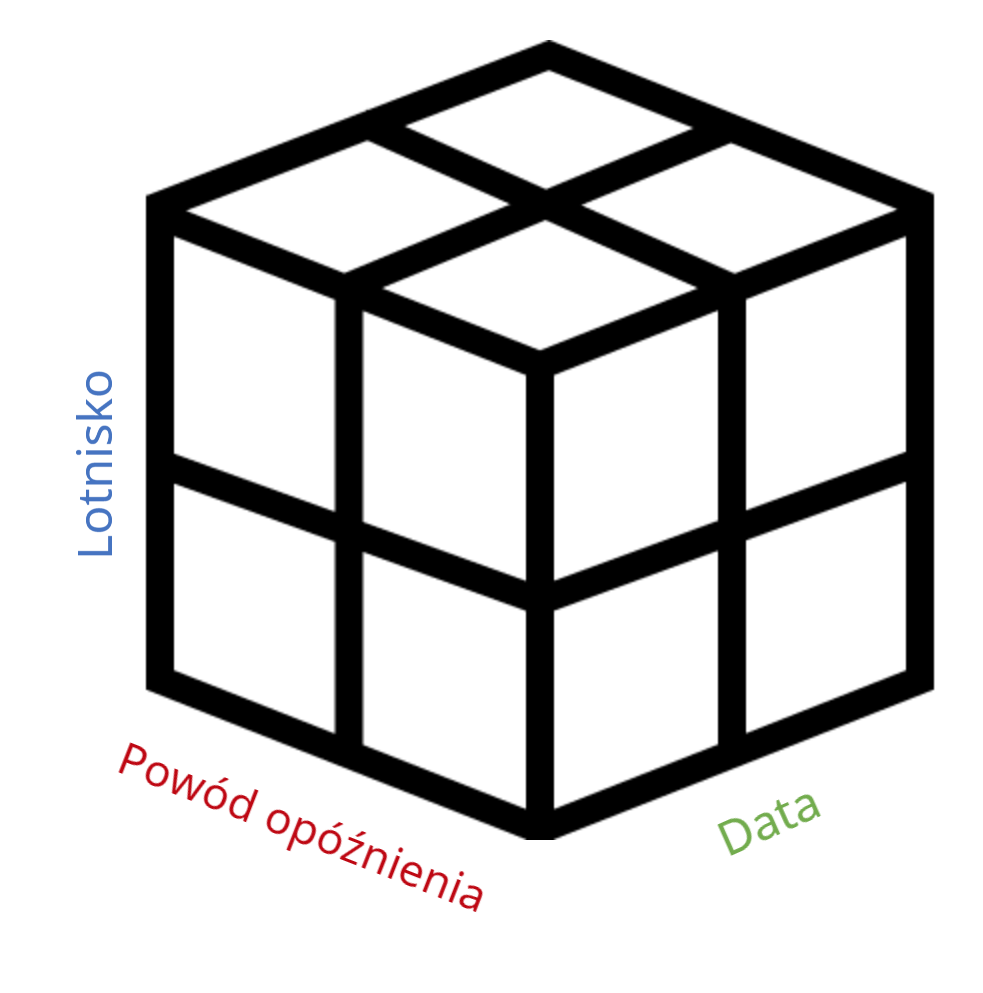
3.2 Agregaty danych

* Średni czas opóźnienia w danym kwartale według linii lotniczej (1)
* Średni czas opóźnienia w danym miesiącu według lotniska (2)
* Łączny koszt lotów według linii lotniczej (3)

1. 3.



2.

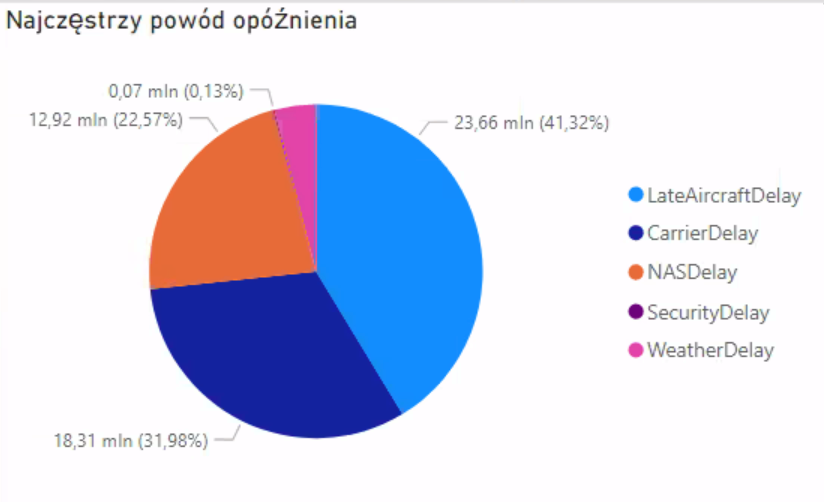


1. Raportowanie
2. Cel Analizy

Celem jest analiza danych dotyczących opóźnień lotów oraz szacowanego kosztu paliwa, na podstawie dostępnych danych powinno być możliwe ustalenia najgorszych kwartałów (w których występuje najwięcej opóźnień), linii lotniczych których średni czas opóźnienia jest największy oraz lotnisk na których występuje największe opóźnienie.

1. Cel przygotowanych wykresów

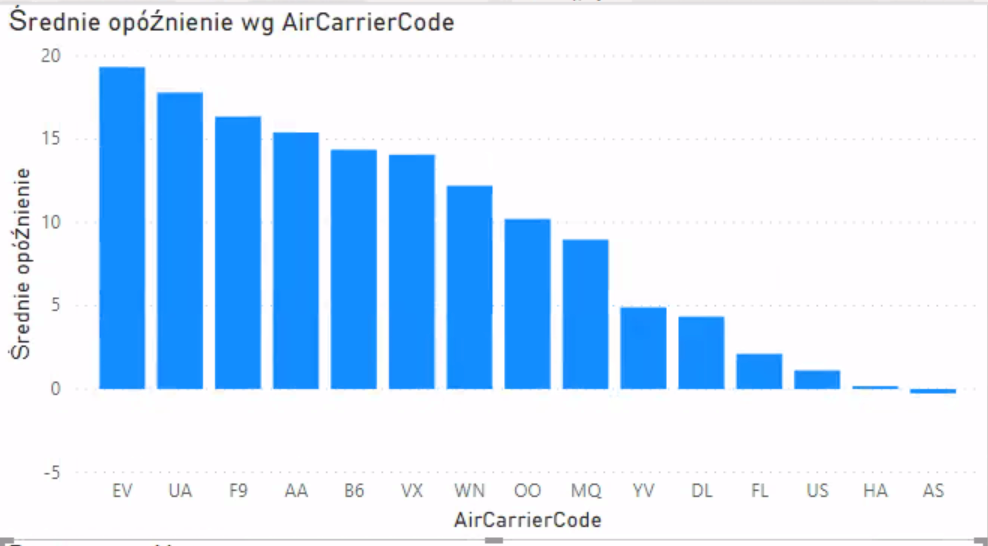
Pierwszym wykresem jest wykres kołowy który przestawia najczęstszy powód odwołania lotów. (agregat 1 lub 2 - suma opóźnień podzielona na kategorie)



Możemy zauważyć że najczęstszym powodem opóźnienia było Late Aircraft Delay, czyli

poprzedni lot tym samym samolotem przyleciał z opóźnieniem, co spowodowało opóźnienie obecnego lotu.

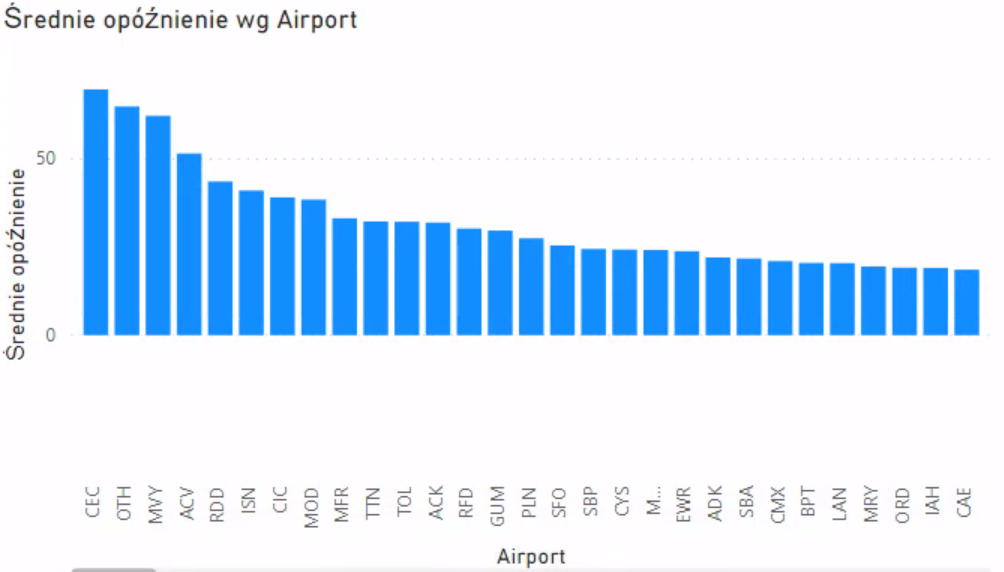
Następnie został utworzony wykres średniego czasu opóźnienia dla każdej z linii lotniczych (agregat 1)



Na wykresie jest widoczny przewoźnik z największą średnią opóźnienia (wszystkie powody) Jest nim ExpressJet.

Kolejnym wykresem jest zestawienie średnich opóźnień występujących na lotniskach

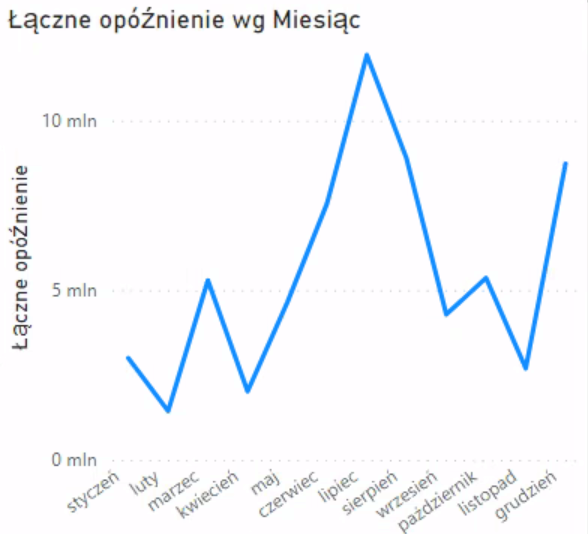
(agregat 2)



Lotnisko, które miało największą średnią opóźnień, to Del Norte County Regional Airport w stanie California.

Kolejnym wykresem analizującym opóźnienia jest wykres łącznego opóźnienia w danym miesiącu

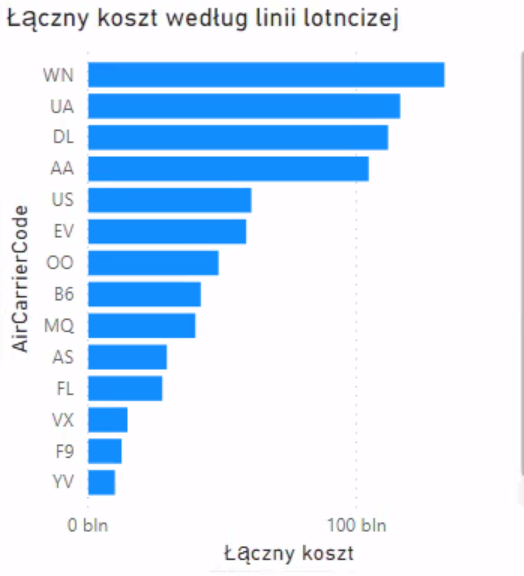
(agregat 2)

.

Możemy zauważyć, że częstotliwość opóźnień jest największa pomiędzy lipcem a wrześniem.

Ostatnim wykresem jest łączny koszt lotów według linii lotniczej, koszt lotów jest traktowany jako współczynnik (dystans \* ceny ropy dla dnia w którym odbywał się lot)

(agregat 3)



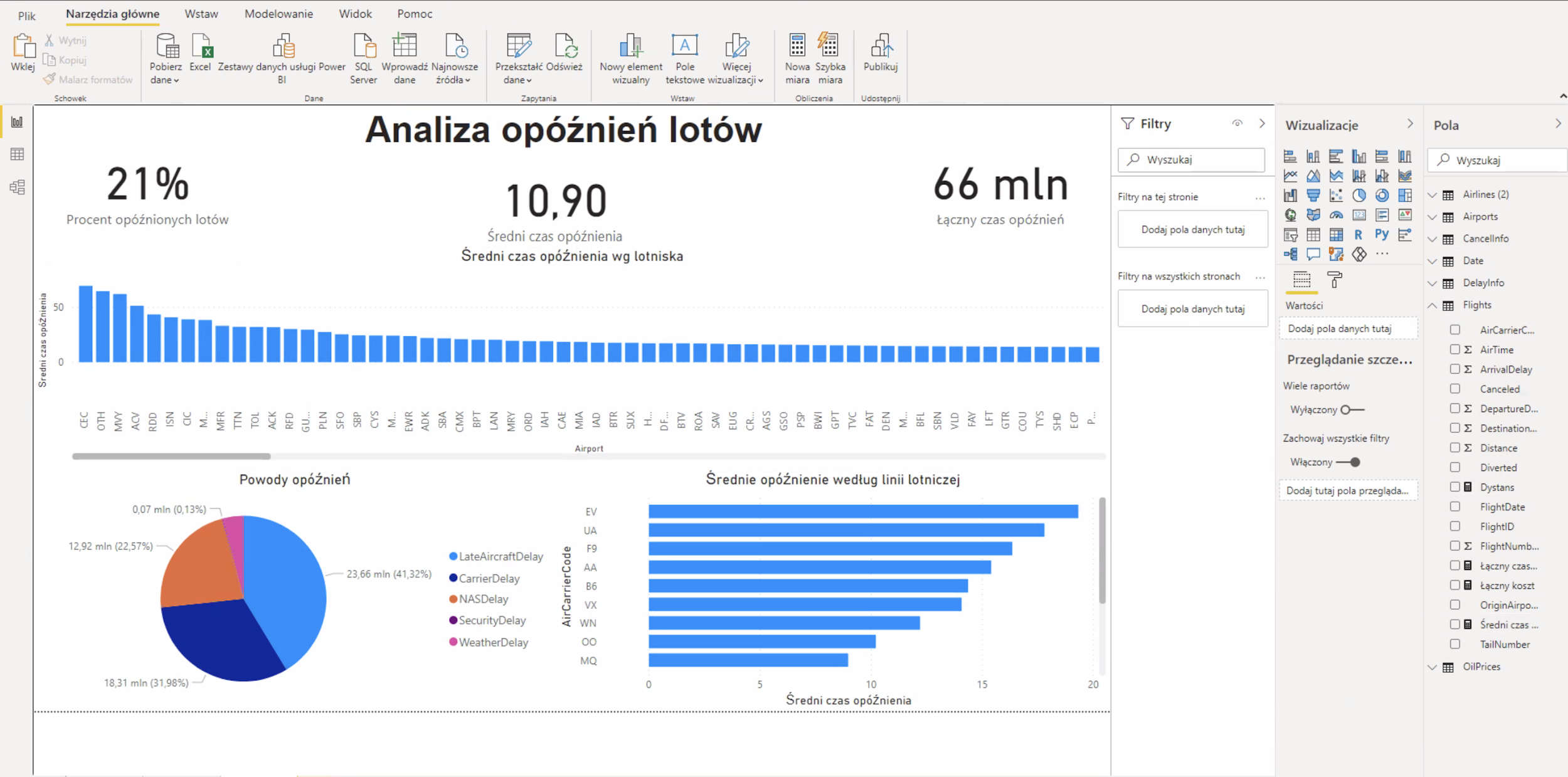
Z wykresu wynika że linią lotniczą która najbardziej wykosztowała się na paliwo jest Southwest Airlines (WN)

*Wszystkie wykresy są dostępne w załączniku w pliku PowerBI*

1. Kokpit menedżerski

Cel - Zaprezentowanie podstawowych informacji na temat opóźnień lotów oraz ich powodów, z perspektywy lotniska oraz perspektywy linii lotniczej. Dane te mogą pomóc potencjalnemu użytkownikowi dowiedzieć się jak statystycznie wyglądają obsługiwane loty przez konkretną linię lotniczą na konkretnym lotnisku.

*Kokpit menedżerski jest dostępny w załączniku jako plik PowerBi*

**