Projekt z przedmiotu Hurtownie Danych i Systemy Business Intelligence

Anna Wawrzyńczak, Izabela Telejko, Grzegorz Zbrzeżny

Maj 2023

Spis treści

1 Wstęp **3**
1.1 Cel biznesowy i wprowadzenie do tematyki projektu . . . 3

2 Opis źródeł danych **4**
2.1 Zbiór Flight reviews . . . 4
2.2 Zbiór Passengers . . . 4
2.3 Zbiory Airlines rankings . . . 5
2.4 Zbiór Aircrafts . . . 5

3 Architektura hurtowni **6**
3.1 Ogólna postać rozwiązania . . . 6
3.2 Diagram . . . 6
3.3 Tabele faktów . . . 6
3.3.1 Tabela FlightReview . . . 6
3.3.2 Tabela AirlineRating . . . 8
3.4 Tabele wymiarów . . . 8
3.4.1 Tabela Passenger . . . 8
3.4.2 Tabela Aircraft . . . 8
3.4.3 Tabela Date . . . 9
3.4.4 Tabela Airline . . . 9
3.4.5 Tabela Geography . . . 9

4 ETL **10**
4.1 Słowniki . . . 10
4.2 Transformacje . . . 10
4.2.1 Tabela DimPassenger . . . 10
4.2.2 Tabela DimAircraft . . . 11
4.2.3 Tabela DimDate . . . 12
4.2.4 Tabela DimGeography . . . 12
4.2.5 Tabela DimAirline . . . 12
4.2.6 Tabela FactAirlineRating . . . 14
4.2.7 Tabela FactFlightReview . . . 14

5 Opis warstwy raportowej **17**
5.1 Przekształcenia danych w PowerBI . . . 17
5.2 Modele danych w Power BI . . . 18
5.2.1 Raport pierwszy . . . 18
5.2.2 Raport drugi . . . 19
5.2.3 Raport trzeci . . . 20
5.3 Spójność danych z warstwą hurtowni danych . . . 21
5.4 Opis raportów . . . 21

6 Podsumowanie biznesowych rezultatów projektu **23**

7 Przeprowadzone testy **24**
7.1 Ogólne testowanie . . . 24
7.2 Data Accuracy . . . 24
7.2.1 DimAirline . . . 24
7.2.2 DimAircraft . . . 24

7.2.3 FactFlightReview . . . 25
7.3 Completeness . . . 25
7.3.1 Braki danych w FlightReviews . . . 25
7.3.2 Test przekształceń danych w wymiarze DimPassenger . . . 25
7.4 Consistency . . . 25
7.4.1 Porównanie liczby wierszy w poszczególnych tabelach hurtowni ze żródłem . . . 25
7.4.2 Porównanie liczby wierszy po złączeniu tabel faktowych z wymiarami . . . 26
7.5 Uniqueness . . . 26
7.5.1 Sprawdzenie wystąpień duplikatów wśród unikalnych kolumn . . . 26
7.6 Validity . . . 26
7.6.1 Zakres miarek dla faktu FlightReview . . . 26
7.6.2 Zakres miarek dla faktu AirlineRating . . . 27
7.6.3 Porównanie ze żródłem sumy miarek dla faktu FlightReview . . . 27
7.6.4 Porównanie ze żródłem sumy miarek dla faktu AirlineRating . . . 27
7.6.5 Testy kolumn kategorycznych . . . 27
7.7 Test SCD2 . . . 28

8 Podział pracy w zespole **29**

Rozdział 1

Wstęp

1.1 Cel biznesowy i wprowadzenie do tematyki projektu

Celem projektu jest analiza opinii pasażerów o odbytych lotach i ich zestawienie z ogólnym rankingiem linii lotniczych stworzonym przez Airhelp. Analiza umożliwi ulepszenie jakości poszczególnych sektorów usług oferowanych przez linie lotnicze, wskazanie ich mocnych i słabych stron, porównanie zadowolenia pasażerów na przestrzeni lat 2013-2019 oraz szukanie zależności między warunkami podróży a poziomem satysfakcji podróżujących.

Na podstawie stworzonych przez nas raportów poszczególne linie lotnicze będą mogły opracować plan zmian i ulepszeń w ich usługach, aby móc zadowolić jak największą liczbę pasażerów, podnieść swoją reputację oraz wyróżnić się wśród konkurentów.

Rozdział 2

Opis źródeł danych

2.1 Zbiór Flight reviews

Zbiór pochodzi z portalu kaggle (link:`https://www.kaggle.com/datasets/efehandanisman/skytrax-airline-reviews`) i zawiera opinie pasażerów odnośnie odbytych lotów z lat 2005-2019. Zbiór ma 16 kolumn i 64 017 rekordów:

1. airline - nazwa linii lotniczej (varchar(100))
2. overall - ogólna ocena lotu od 1 do 10 (int)
3. author - imię i nazwisko oceniającego (varchar(100))
4. review_date - data wystawienia opinii (varchar(100))
5. aircraft - model samolotu (varchar(100))
6. traveller_type - cel podróży pasażera (varchar(50))
7. cabin - klasa, którą leciał recenzent (varchar(50))
8. route - trasa lotu (varchar(200))
9. date_flown - data lotu (varchar(100))
10. seat_comfort - ocena komfortu siedzeń od 1 do 5 (int)
11. cabin_service - ocena obsługi w samolocie od 1 do 5 (int)
12. food_bev - ocena jedzenia i napojów od 1 do 5 (int)
13. entertainment - ocena zapewnionej rozrywki podczas lotu od 1 do 5 (int)
14. ground_service - ocena osługi na lotnisku od 1 do 5 (int)
15. value_for_money - ocena oferownanej jakości za cenę lotu od 1 do 5 (int)
16. recommended - czy oceniający poleciłby dany lot (”yes”/”no”/”/N”) (varchar(10))

Z uwagi na stosunkowo niewielką liczbę recenzji z lat 2005-2012 w naszej hurtowni będziemy składować tylko dane od roku 2013. Ze znajdujących się w zbiorze kolumn nie będziemy brać kolumny date_flown, a zostawimy jedynie review_date. Wartości kolumny author zostaną zamienione na wygenerowane syntetycznie nowe imiona i nazwiska, ponieważ zbiór o pasażerach linii lotniczych jest generowany syntetycznie.

2.2 Zbiór Passengers

Do pełnego rozwiązania potrzebne było sztuczne wygenerowanie niektórych danych. Za pomocą generatora online stworzono zbiór Passengers, który przedstawiał informacje na temat pasażerów linii lotniczych. W tym zbiorze zawarte są dane zarówno o podróżujących, którzy wystawili co najmniej jedną opinię w latach 2013 - 2019 jak i o użytkownikach, którzy odbyli przynajmniej jeden lot i nie wystawili opinii w rozważanym przedziale czasowym. Zbiór Passengers ma 19 724 rekordów i następujące kolumny:

1. id - numer identyfikacyjny pasażera (int)

2. firstname - pierwsze imię pasażera (varchar(50))
3. lastname - nazwisko pasażera (varchar(50))
4. Profession - zawód podróżującego (varchar(100))
5. Age - wiek pasażera (int)
6. Status - status podróżującego; w przypadku wykupienia statusu Gold lub Silver użytownikowi zapewniony jest pakiet zniżek i udogdnień (varchar(50))
7. FlightsCountLastYear - liczba lotów, które odbył podróżujący w ciągu poprzedniego roku.

2.3 Zbiory Airlines rankings

Od 8 lat co najmniej raz do roku Airhelp publikuje globalny ranking linii lotniczych (dostępny na stronie `https://www.airhelp.com/en-int/airhelp-score/airline-ranking/`). Przy użyciu danych z podanej strony oraz zbioru znalezionego na platformie dataworld, który zawiera zescrappowane dane z rankingu Airhelp z 2018 roku (link: `https://data.world/dataremixed/2018-airline-rankings-by-airhelp/activity`) zostały stworzone zbiory airlines rankings, z których każdy opisuje ranking linii lotniczych w danym roku z przedziału 2015-2019. Ze względu na dużą liczbę opinii o lotach w latach 2013 i 2014 wygenerowano sztucznie dwa nowe rankingi dla linii lotniczych, które znajdowały się w rankingu z 2015 r. Zbiory zawierają różną liczbę rekordów. W rankingu z 2013 znajduje się 41 wierszy, zaś w zbiorach z lat 2018 i 2019 są 72 rekordy. Przy złączeniu wszystkich zbiorów wychodzi razem 404 rekordy. W każdym ze zbiorów z Airlines rankings znajdują się następujące kolumny:

1. rank - pozycja w rankingu (int)
2. airline_name - nazwa linii lotniczej (varchar(100))
3. Punctuality - ogólna ocena punktuaności lotu (int)
4. Service_Quality - ogólna ocena jakości serwisu w trakcie lotu (int)
5. Claim_processing - ogólna ocena szybkości weryfikacji reklamacji i przetwarzania rekompensat (int)
6. Airhelp_score - średnia z trzech powyżej opisanych ocen (int)
7. Country - kraj, w który znajduje się główna siedziba danej linii lotniczej. (varchar(100))

Ogólne oceny zostały obliczone na podstawie średniej z ocen pasażerów o danym sektorze dla każdej z linii lotniczych. W trakcie tworzenia modelu hurtowni usuniemy kolumnę Airhelp_score ze względu na to, że jest to średnia z innych ocen a przeprowadzana analiza będzie głównie opierać się na podstawie wartości z Punctuality, Service_Quality oraz Claim_processing.

2.4 Zbiór Aircrafts

Zbiór Aircrafts zawiera informacje o modelach samolotów (źródło:`http://www.lsv.fr/~sirangel/teaching/dataset/index.html`). Dane zawierają 218 wierszy i poniżej opisane kolumny:

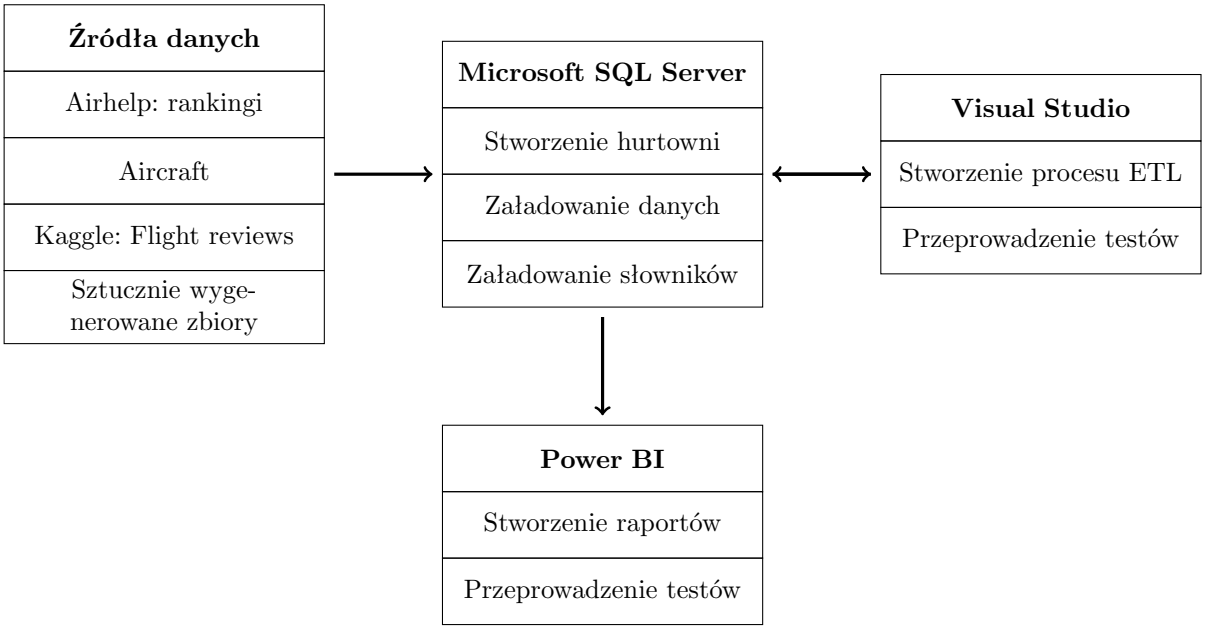
1. Name - nazwa modelu samolotu (varchar(100))
2. ICAO - skrót ICAO nazwy modelu (varchar(50))
3. IATA - skrót IATA modelu (varchar(50))
4. Capacity - liczba miejsc siedzących w samolocie (bigint)
5. Country - miejsce produkcji modelu (varchar(100))

W modelu nie będziemy korzystać z kolumny zawierającej skrót IATA, wystarczająca będzie kolumna ze skrótem ICAO. Ponadto, zakładamy unikalność kolumn z nazwą modelu.

Rozdział 3

Architektura hurtowni

3.1 Ogólna postać rozwiązania



3.2 Diagram

Nasza hurtownia składa się z 7 tabel, w tym 2 faktowych. Połączone one są w schemat galaktyki, co przedstawia Rysunek 3.1. Do każdej tabeli został dodany unikalny automatycznie inkrementujący się klucz główny.

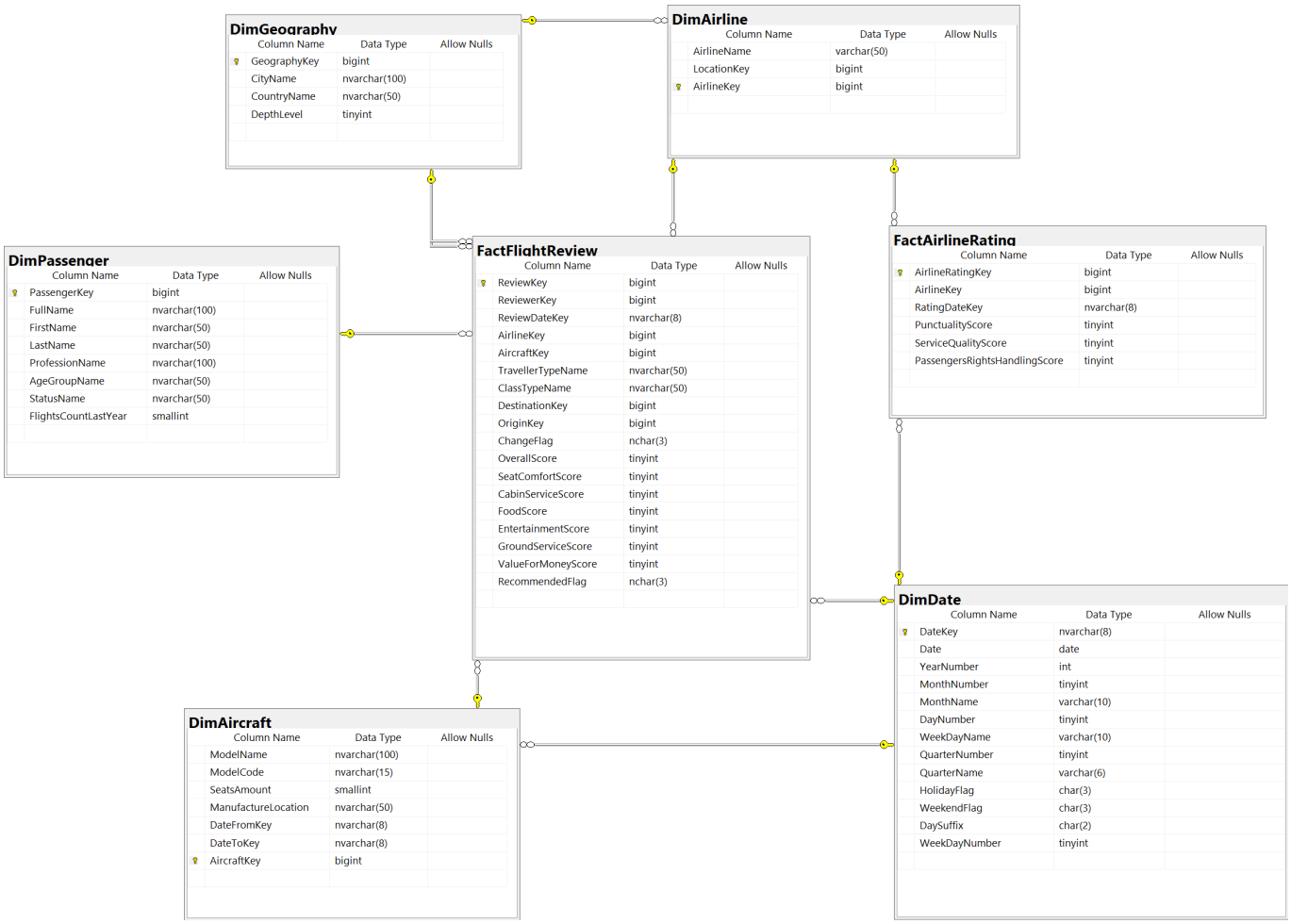
3.3 Tabele faktów

W przedstawianym modelu hurtowni wyróżniamy dwie tabele faktów: tabela FlightReview oraz tabela AirlineRating.

3.3.1 Tabela FlightReview

Pierwszym faktem w naszym modelu będą opinie o lotach wystawione przez pasażerów. Tabela powstaje na podstawie zbioru danych Flight reviews. Kolumny jakie się w niej znajdują to:

1. ReviewKey - unikalne ID faktu (klucz główny)
2. ReviewerKey - ID pasażera (klucz obcy - relacja do wymiaru pasażera)
3. ReviewDateKey - ID daty wystawienia opinii (klucz obcy - relacja do wymiaru daty)
4. AirlineKey - ID linii lotniczej (klucz obcy - relacja do wymiaru linii lotniczych)



Rysunek 3.1: Diagram modelu hurtowni danych

5. AircraftKey - ID modelu samolotu (klucz obcy - relacja do wymairu modeli samolotów)
6. TravelerTypeName - nazwa typu podróżującego (atrybut)
7. ClassTypeName - nazwa klasy, którą leciał podróżujący (atrybut)
8. DestinationKey - ID miejsca docelowego lotu (klucz obcy - relacja do wymiaru geograficznego)
9. OriginKey - ID miejsca wylotu (klucz obcy - realcja do wymiaru geograficznego)
10. ChangeFlag - informacja czy podróż była z przesiadką (atrybut)
11. OverallScore - ogólna ocena lotu od 1 do 10 (miarka)
12. SeatComfortScore - ocena komfortu siedzeń od 1 do 5 (miarka)
13. CabinServiceScore - ocena obsługi w samolocie od 1 do 5 (miarka)
14. FoodScore - ocena jedzenia i napojów od 1 do 5 (miarka)
15. EntertainmentScore - ocena zapewnionej rozrywki podczas lotu od 1 do 5 (miarka)
16. GroundServiceScore - ocena obsługi na lotnisku od 1 do 5 (miarka)
17. ValueForMoneyScore - ocena oferowanej jakości za cenę lotu od 1 do 5 (miarka)
18. RecommendedFlag - czy oceniający poleciłby dany lot (miarka)

Z uwagi na to, iż w zbiorze Flight reviews nie ma żadnego innego identyfikatora pasażera wystawiającego opinię, nie jesteśmy w stanie rozróżnić osób o identycznych imionach i nazwiskach. Takie osoby będziemy w naszym modelu zatem traktować jako jedną (sytuacji takich jednak nie powinno być wiele, ponieważ wszystkich pasażerów jest

łącznie 19 724 - szanse na powtarzające się godności są niewielkie). Tabela będzie aktualizowana raz na tydzień, aby móc mieć dostęp do w miarę aktualnych danych. Będą dodawane do niej nowe rekordy, a raz wstawione nie będą aktualizowane. Wynikowa tabela powinna mieć 18 kolumn i 64017 wierszy.

3.3.2 Tabela AirlineRating

Drugim faktem będą oceny poszczególnych sektorów dla linii lotniczej w danym roku. Tabela faktowa powstanie na podstawie zbiorów: Airlines_ratings_2013-Airlines_ratings_2019 i Aircrafts. Klucz główny jest tworzony jako konkatenacja RatingYear i AirlineKey. Kolumny jakie się będą w niej znajdować to:

1. AirlineRatingKey - unikalne ID faktu (klucz główny)
2. AirlineKey - ID linii lotniczej (klucz obcy - relacja do wymiaru linii lotniczych)
3. RatingDateKey - ID daty, dla której jest ocena (klucz obcy do wymiaru daty)
4. PunctualityScore - ocena punktualności lotu od 1 do 10 (miarka)
5. ServiceQualityScore - ocena jakości serwisu w trakcie lotu od 1 do 10 (miarka)
6. PassengersRightsHandlingScore - ocena szybkości weryfikacji reklamacji i przetwarzania rekompensat od 1 do 10 (miarka)

Dla tych danych mamy podany jedynie rok w którym wystawiono ocenę tak więc aby była możliwość ustawienia relacji do wymiaru daty, ustalamy wartość RatingDateKey na pierwszego stycznia tego roku. Dane będą aktualizowane raz do roku, gdy dostępny będzie nowy zbiór z ocenami za kolejny rok. Wówczas do tabeli dodane zostaną nowe rekordy z danym rokiem, a raz dodane rekordy nie będą aktualizowane. Wynikowa tabela powinna mieć 6 kolumny i 404 wierszy.

3.4 Tabele wymiarów

3.4.1 Tabela Passenger

Jednym z wymiarów tabeli faktów FlightReview jest tabela Passenger, które zawiera informacje na temat pasażerów linii lotniczych, wzięte ze zbioru Passengers. W Passenger będą zawarte następujące kolumny:

1. PassengerKey - unikalne ID pasażera (klucz główny oraz klucz obcy - relacja do FlightReview)
2. FullName - imię i nazwisko pasażera
3. FirstName - imię podróżującego
4. ProfessionName - nazwa zawodu pasażera
5. AgeGroupName - zmienna kategoryczna stworzona na podstawie danych o wieku podróżujących
6. StatusName - status podróżującego
7. FlightsCountLastYear - liczba lotów odbytych przez pasażera

Tabela Passengers będzie aktualizowana raz na tydzień. Zakładamy unikalność tabeli ze względu na FullName pasażera. Wynikowa tabela powinna mieć 7 kolumn i 19 724 wierszy.

3.4.2 Tabela Aircraft

Tabela DimAircraft jest wymiarem SCD typu 2 dedykowanym dla faktu FlightReview. Powstaje na podstawie zbioru Aircrafts. Jej kolumny to:

1. AircraftKey - unikalne ID modelu samolotu (klucz główny)
2. ModelName - nazwa modelu
3. ModelCode - skrót nazwy modelu
4. SeatsAmount - liczba siedzeń w modelu samolotu
5. ManufactureLocation - miejsce produkcji modelu

6. DateFromKey - data załadowania rekordu (klucz obcy do tabeli daty)
7. DateToKey - data wskazująca do kiedy rekord obowiązuje (klucz obcy do tabeli daty)

Tabela będzie aktualizowana raz na kwartał, ponieważ zanim pasażerowie będą mogli lecieć nowym modelem samolotu upłynie jeszcze sporo czasu - model będzie testowany. Nie są dla nas ważne zatem dane z ostatnich paru miesięcy i możemy sobie pozwolić na rzadszą aktualizację. Do tabeli będziemy dodawać nowe rekordy (nowe modele) oraz aktualizować już się w niej znajdujące, jeśli parametry pewnego modelu się zmienią. Zakładamy unikalność kolumny ModelName. Wynikowa tabela powinna mieć 7 kolumn i 218 wierszy.

3.4.3 Tabela Date

W modelu znajduje się również wymiar daty, do którego odnoszą się wszystkie kolumny zawierające informację o pewnej dacie. Kluczem głównym tabeli będzie numer tworzony przez konkatenację roku, miesiąca i dnia. Dane w tabeli nie będą periodycznie aktualizowane. Jej kolumny to:

1. DateKey - unikalne ID daty (klucz główny)
2. Date - data
3. YearNumber - numer roku
4. MonthNumber - numer miesiąca
5. MonthName - nazwa miesiąca
6. DayNumber - dzień miesiąca
7. WeekDayName - nazwa dnia tygodnia
8. QuarterNumber - numer kwartału
9. QuarterName - nazwa kwartału
10. HolidayFlag - czy danego dnia jest święto
11. WeekendFlag - czy danego dnia jest weekend
12. DaySuffix - skrót nazwy dnia tygodnia
13. WeekDayNumber - numer dnia w tygodniu.

3.4.4 Tabela Airline

Tabela ta zawiera dokładniejsze informacje na temat linii lotniczych, którym strona AirHelp nadała rankingi. Tabela powstaje na podstawie zbiorów: Airlines_ratings_2013-Airlines_ratings_2019. Aktualizacja danych w tej tabeli odbywać się będzie raz na rok, gdy AirHelp wypuści nowy ranking. Jej kolumny to:

1. AirlineKey - unikalne ID wymiaru (klucz główny)
2. AirlineName - pełna nazwa linii lotniczej
3. LocationKey - ID kraju, w którym znajduje się siedziba główna linii lotniczej (klucz obcy do tabeli geografii)

Ponadto, zakładamy unikalność kolumny AirlineName. Wynikowa tabela powinna mieć 3 kolumny i 117 wierszy.

3.4.5 Tabela Geography

Jest to outrigger podłączony do tabel, które zawierają w sobie jakąś lokalizację. Wymiar ten opisuje dokładniej poszczególne lokalizacje. Dane w tej tabeli nie są periodycznie aktualizowane. Kolumny w tej tabeli to:

1. GeographyKey - unikalne ID wymiaru (klucz główny)
2. CountryName - nazwa kraju
3. CityName - nazwa miasta
4. DepthLevel - wartość wskazującą czy rekord dotyczy tylko kraju czy kraju i miasta.

DepthLevel przyjmuje wartość 1, gdy rekord dotyczy tylko kraju a informacja o mieście nie jest podana (w tabeli nazwa miasta jest zastępywana przez Unknown). Jeżeli DepthLevel jest równe 2 to rekord dotyczy miasta i kraju i obie te wartości są znane.

Rozdział 4

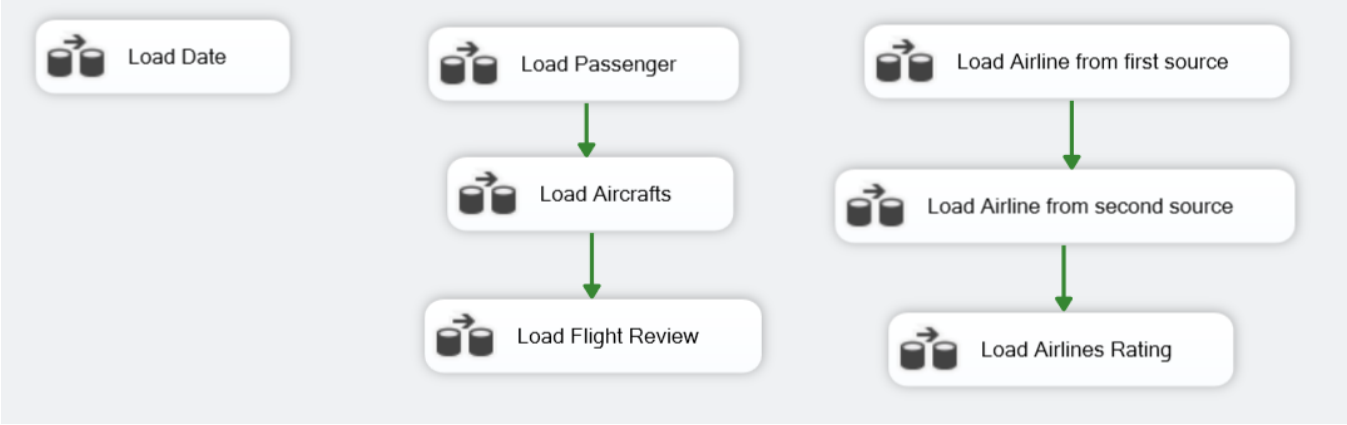
ETL

4.1 Słowniki

Do niektórych transformacji w zakresie mapowania potrzebne było stworzenie słowników. Słowniki MonthsMappingDictionary, AircraftMappingDictionary, AirlinesMappingDictionary, CitiesMappingDictionary oraz CountriesMappingDictionary zostały stworzone w Jupyter Notebook lub MS SQL Server. Słownik AircraftMappingDictionary został wykorzystany, aby mapować nazwy i skróty modeli samolotów podanych przez recenzentów w Flights reviews na odpowiednie nazwy samolotów z źródła danych Aircraft. W analogiczny sposób stworzono słownik AirlinesMappingDictionary do ujednolicenia nazw linii lotniczych pomiędzy liniami lotniczymi w recenzjach a liniami w rankingach. Słowniki CitiesMappingDictionary i CountriesMappingDictionary zostały stworzone w Jupyter Notebook przy wykorzystaniu biblioteki geopy.geocoders. Za pomocą tej biblioteki na podstawie miasta lub skrótu miasta podanego przez recenzenta (z kolumny route) przypisano nazwę kraju, w którym znajduje się miasto. Słownik CountriesMappingDictionary pozwala na ujednolicenie nazw krajów, tak aby można było je porównać z krajami w tabeli geografii. Słownik MonthsMappingDictionary został stworzony, aby zamieniać nazwy miesięcy na numer miesiąca w roku.

4.2 Transformacje

Ładowanie danych dla każdej tabeli polegało na wczytaniu plików csv. Ponadto rekordy, które wywołały błędy, będą zapisywane jako pliki tekstowe. Cały process ETL został pokazany na Rysunku 4.1.
Poniżej przedstawiono dokładny opis transformacji wykonanych w obrębie poszczególnych tabel.

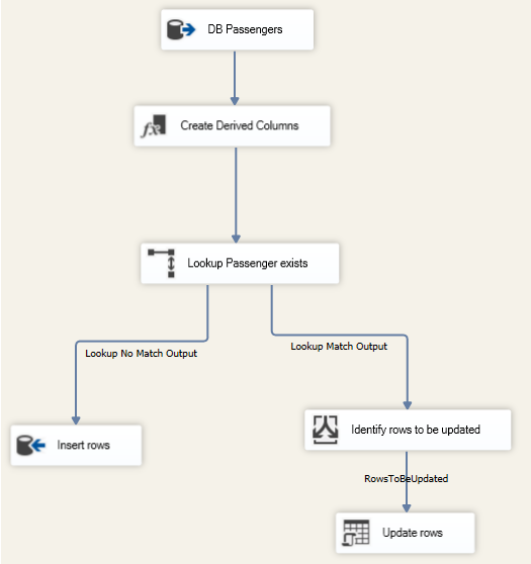


Rysunek 4.1: Control flow projektu

4.2.1 Tabela DimPassenger

Opis transformacji (przedstawiony na Rysunku 4.2):

1. Create Derived Columns - dodanie kolumny AgeGroup, która dzieli pasażerów względem wieku na 4 grupy: Young Adult (poniżej 25 roku życia), Adult (poniżej 44 roku życia), Middle-age (poniżej 60 roku życia) i Ederly (powyżej 60 lat). Ponadto, dodano kolumnę FullName, która została stworzona z połączenia kolumn FirstName i LastName; Dodatkowo sprecyzowano typ danych w nowych kolumnach jako nvarchar lub int.



Rysunek 4.2: Data flow dla wymiaru Passenger

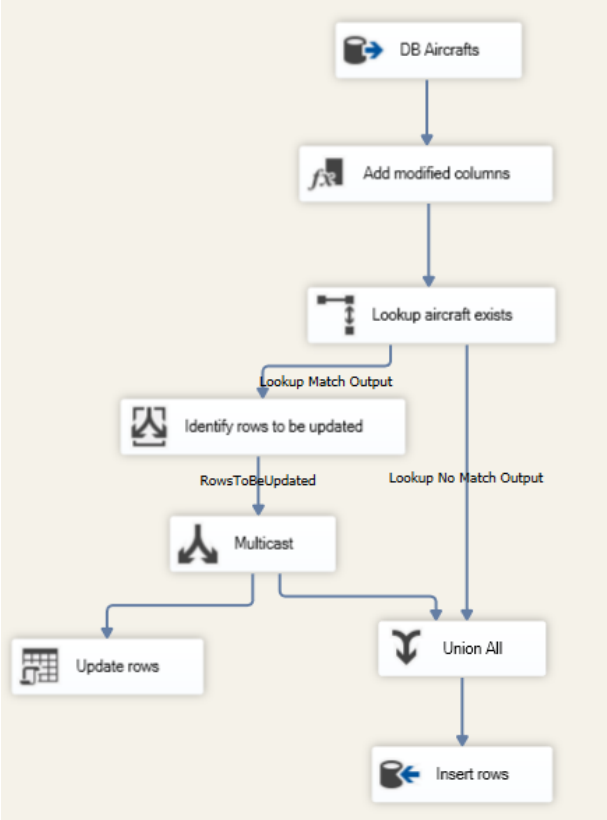
2. Lookup exists, Insert, Update - sprawdzenie, czy dany rekord o recenzencie już istnieje w hurtowni i konsekwentnie dodanie go do hurtowni lub zaktualizowanie informacji o danym recenzencie.

4.2.2 Tabela DimAircraft

Tabela Aircraft została załadowana w sposób SCD typu 2, stąd zostały dodane kolumny DateFromKey i DateToKey.

Opis transformacji (przedstawiony na Rysunku 4.3):

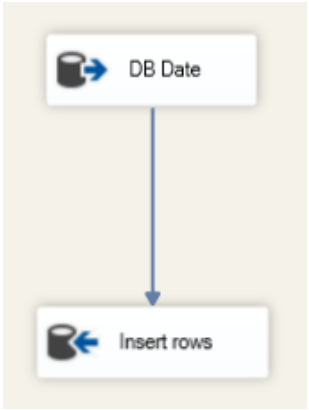
1. Add modified columns - dodanie kolumn ModelName, ModelCode i ManufactureLocation po usunięciu cudzysłowów z orginalnych kolumn (dodatkowo braki danych w kolumnie ICAO zostały zastąpione przez Unknown); dodanie kolumny SeatsAmount na podstawie wartości w Capacity przy czym braki danych zostały zastąpione przez -1; dodanie kolumny DateFromKey, która powstała na podstawie daty wczytania rekordu; dodanie DateToKey jako 99991231.
   Dodatkowo sprecyzowano typ danych w nowych kolumnach jako nvarchar lub int.
2. Lookup exists - sprawdza czy istnieje już dany rekord;
3. Identify rows to be updated - sprawdza w przypadku istnienia rekordu nr klucza głównego tego rekordu;
4. Multicast;
5. Update - zmiana DateToKey z 99991231 na datę zmiany wiersza;
6. Union All - łączy zarówno nowe rekordy jak i zaktualizowane rekordy i dodaje je jako nowe wiersze
7. Insert.



Rysunek 4.3: Data flow dla wymiaru Aircraft

4.2.3 Tabela DimDate

DimDate został załadowany na podstawie przykładowego gotowego wymiaru daty przedstawionego na zajęciach. Transformacje są przedstawione na Rysunku 4.4.



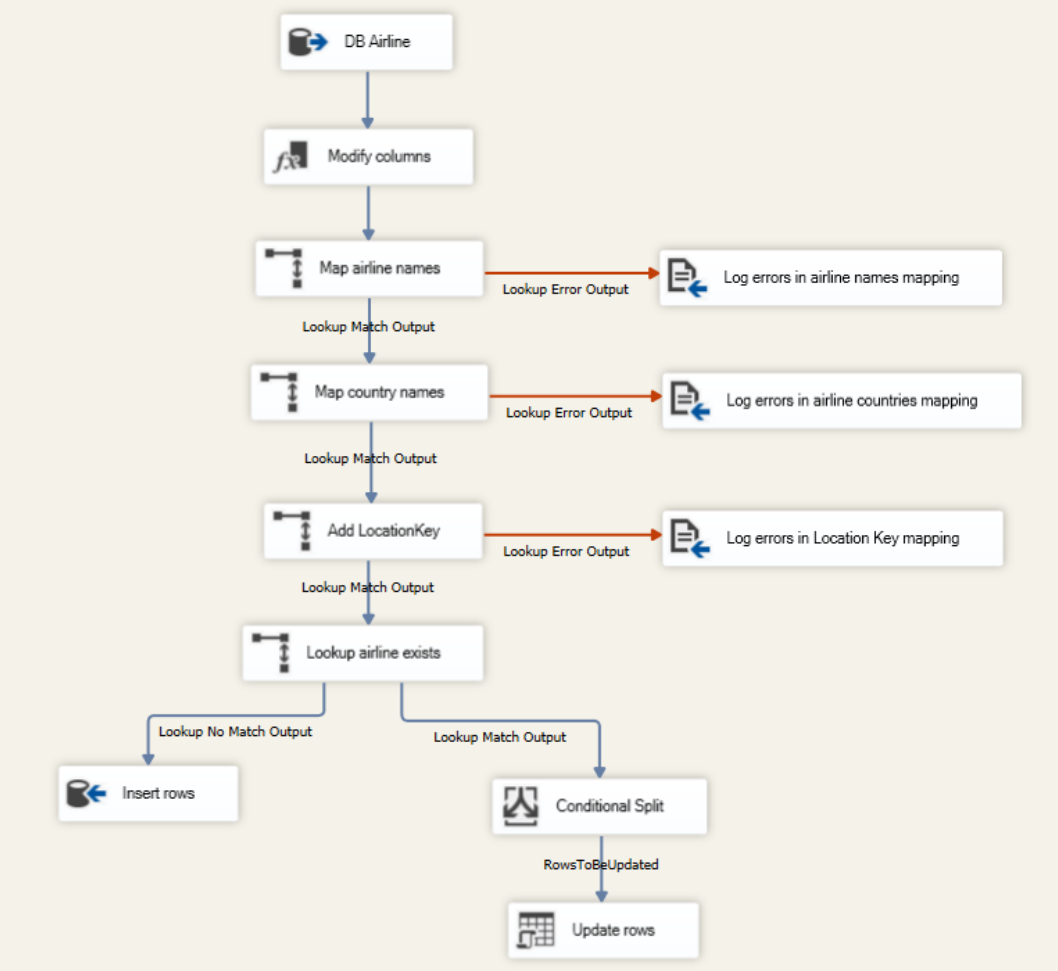
Rysunek 4.4: Data flow dla wymiaru Date

4.2.4 Tabela DimGeography

Tabela DimGeography powstała na podstawie dwóch słowników CitiesMappingDictionary i CountiresMappingDictionary. Po załadowaniu obu słowników do hurtowni złączono je w jedną tabelę. Tabela ta zawiera wszystkie zmapowane miasta z krajami, w których się znajdują oraz poziomem ziarnistości 2 a także kraje pochodzenia linii lotniczych z poziomem ziarnistości 1.

4.2.5 Tabela DimAirline

Opis transformacji dla danych z pierwszego źródła czyli Airline rankings (przedstawiony na Rysunku 4.5):



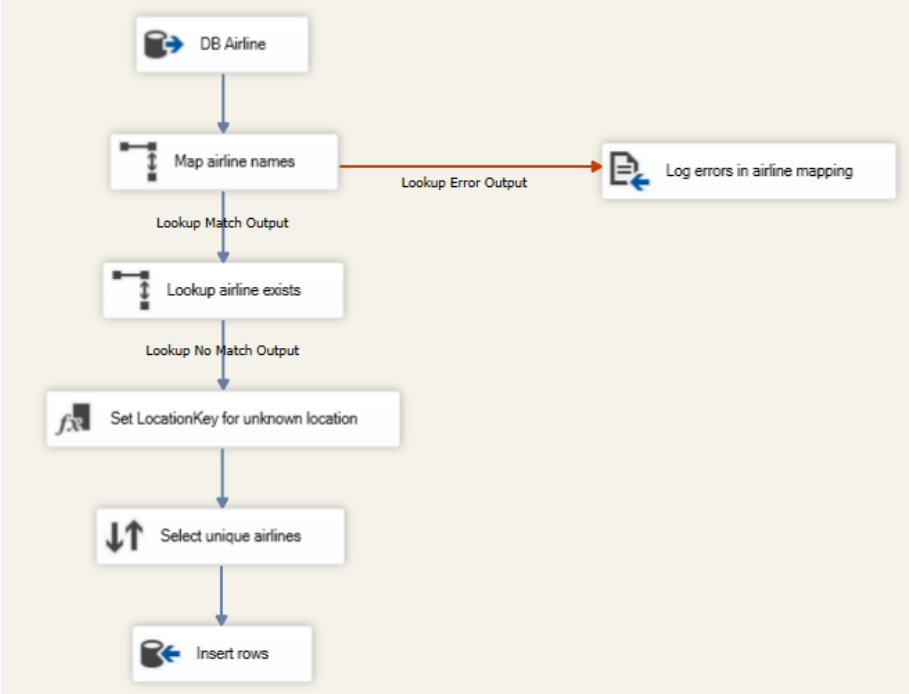
Rysunek 4.5: Data flow dla wymiaru Airline dla danych z pierwszego źródła

1. Modify columns - dodanie kolumny DepthLevel z wartością 1 (główne siedziby linii lotniczych mamy podane w źródle tylko jako państwa); dodanie kolumn Country oraz airline_name po usunięciu cudzysłowów z orginalnych kolumn;
   Dodatkowo sprecyzowano typ danych w nowych kolumnach jako nvarchar lub int.
2. Map airline names - mapowanie airline names przy wykorzystaniu słownika AirlinesMappingDictionary;
3. Map country names i add LocationKey - mapowanie country names przy wykorzystaniu słownika CountriesMappingDictionary i dodawanie kolumny LocationKey;
4. Lookup exists, Insert, Update.

Opis transformacji dla danych z drugiego źródła czyli FlightReviews (przedstawiony na Rysunku 4.6):

1. Map airline names - mapowanie airline names przy wykorzystaniu słownika AirlinesMappingDictionary;
2. Lookup exists;
3. Set LocationKey for unknown location - ustawianie kraju pochodzenia linii lotniczych jako unknown dla tych linii lotniczych, których nie ma w rankingach, wtedy również ustawienie LocationKey na 0;
4. Select unique airlines - wybieranie tylko różnych nazw linii lotniczych z opinii;
5. Insert.

Błędy wynikające z mapowania są zapisywane jako pliki tekstowe.



Rysunek 4.6: Data flow dla wymiaru Airline dla danych z drugiego źródła

4.2.6 Tabela FactAirlineRating

Opis transformacji (przedstawiony na Rysunku 4.7):

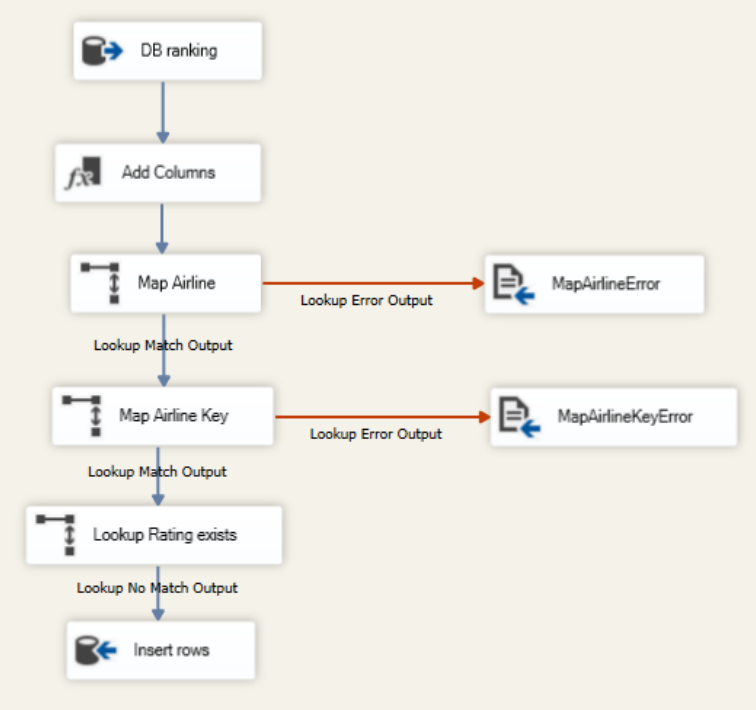
1. Add columns - dodanie RatingDateKey jako 1 stycznia danego roku, z którego pochodzi ranking; usuniecie z nazwy linii lotniczej cudzysłowów;
2. Map Airline - mapowanie airline names przy wykorzystaniu słownika AirlinesMappingDictionary;
3. Map Airline Key - dodanie AirlineKey na podstawie AirlineName z tabeli wymiaru Airline;
4. Lookup exists;
5. Insert;

Błędy wynikające z mapowania są zapisywane jako pliki tekstowe.

4.2.7 Tabela FactFlightReview

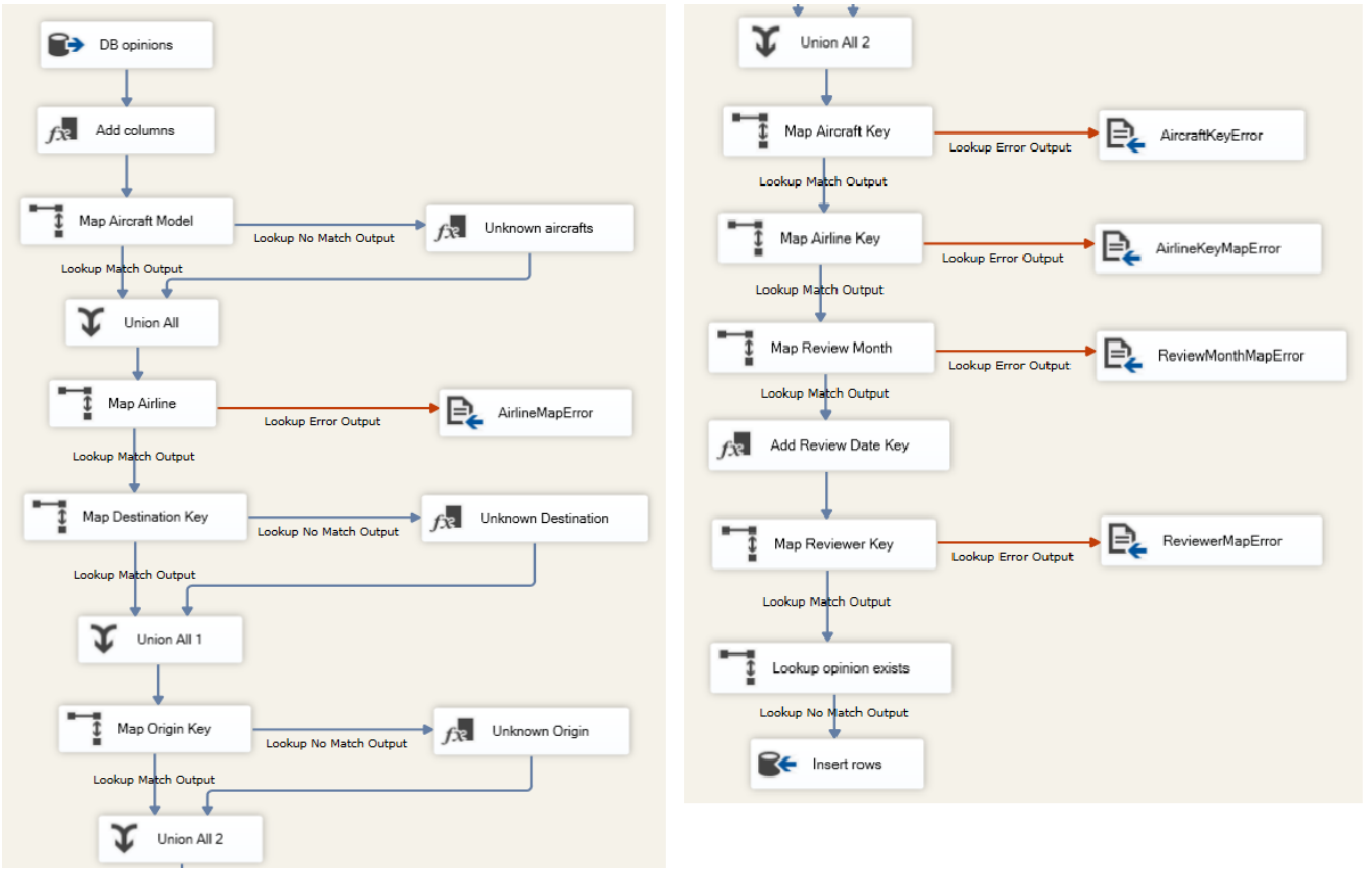
Opis transformacji (przedstawiony na Rysunku 4.8):

1. Add columns - dodanie kolumny ChangeFlag na podstawie obecności 'via' w kolumnie route z źródła danych; dodanie kolumn OriginCity na podstawie wyciągniętej nazwy miasta wylotu i DestinationCity na podstawie wyciągniętej nazwy miasta przylotu z kolumny route, nazwy zostały wyciągnięte bazując na schemacie w jakim są przedstawione rekordy w route, tzn. "... to ... via ..." ; dodanie ReviewMonthName, ReviewDayNumber i ReviewYearNumber na podstawie odpowiednich substringow z kolumny review_date; dodanie kolumny DepthLevel, której wartości są ustawiane na 2 (ponieważ mamy informacje o miastach a nie krajach); dodanie kolumny EndDateKey jako 99991231; dodanie kolumny RecommendedFlag na podstawie wartości yes i no w kolumnie Recommended, przy czym braki danych sa zmieniane na Unknown; dodanie kolumny ClassTypeName na podstawie kolumny ClassType, przy czym puste stringi są zastępowane przez Unknown; dodanie kolumny TravellerTypeName na podstawie orginalnej kolumny TravellerType, przy czym puste stringi są zamieniane na Unknown;
2. Map Aircraft Model i Unknown aircrafts - mapowanie nazw modeli samolotów za pomocą słownika AircraftMappingDictionary; nazwy modeli, których nie ma w źródle danych Aircraft czyli te które się nie zmapowały, zostają zamienione na Unknown
3. Union All - łączenie zmapowanych i niezmapowanych nazw modeli samolotów;



Rysunek 4.7: Data flow dla faktu AirlineRating

4. Map Airline - zmapowanie nazw linii lotniczych przy pomocy AirlineMappingDictionary;
5. Map Destination Key i Unknown Destination - zmapowanie miast i krajów za pomocą CitiesMappingDictionary; miasta, których nie udało się zmapować są zamieniane na Unknown, a kraj do nich przypisywany to również Unknown;
6. Union All 1 - łączenie zmapowanych i niezmapowanych miast i krajów dla miasta przylotu;
7. Map Origin Key, Unknown Origin, Union All 2 - analogiczne mapowanie i łączenie jak w poprzednim korku dla DestinationCity;
8. Map Aircraft Key - przypisywanie AircraftKey na podstawie wymiaru DimAircraft;
9. Map Airline Key - przypisywanie AirlineKey na podstawie wymiaru DimAirline;
10. Map Review Month - przy wykorzystaniu słownika MonthsMappingDictionary nazwy miesięcy zostały zamienione na numer miesiąca w roku;
11. Add Review Date Key - dodanie ReviewDateKey przy wykorzystaniu DimDate;
12. Map Reviewer Key - dodawanie ReviewerKey na podstawie mapowania imienia i nazwiska recenzenta z DimPasenger;
13. Lookup exists;
14. Insert.



Rysunek 4.8: Data flow dla faktu FlightReview

Rozdział 5

Opis warstwy raportowej

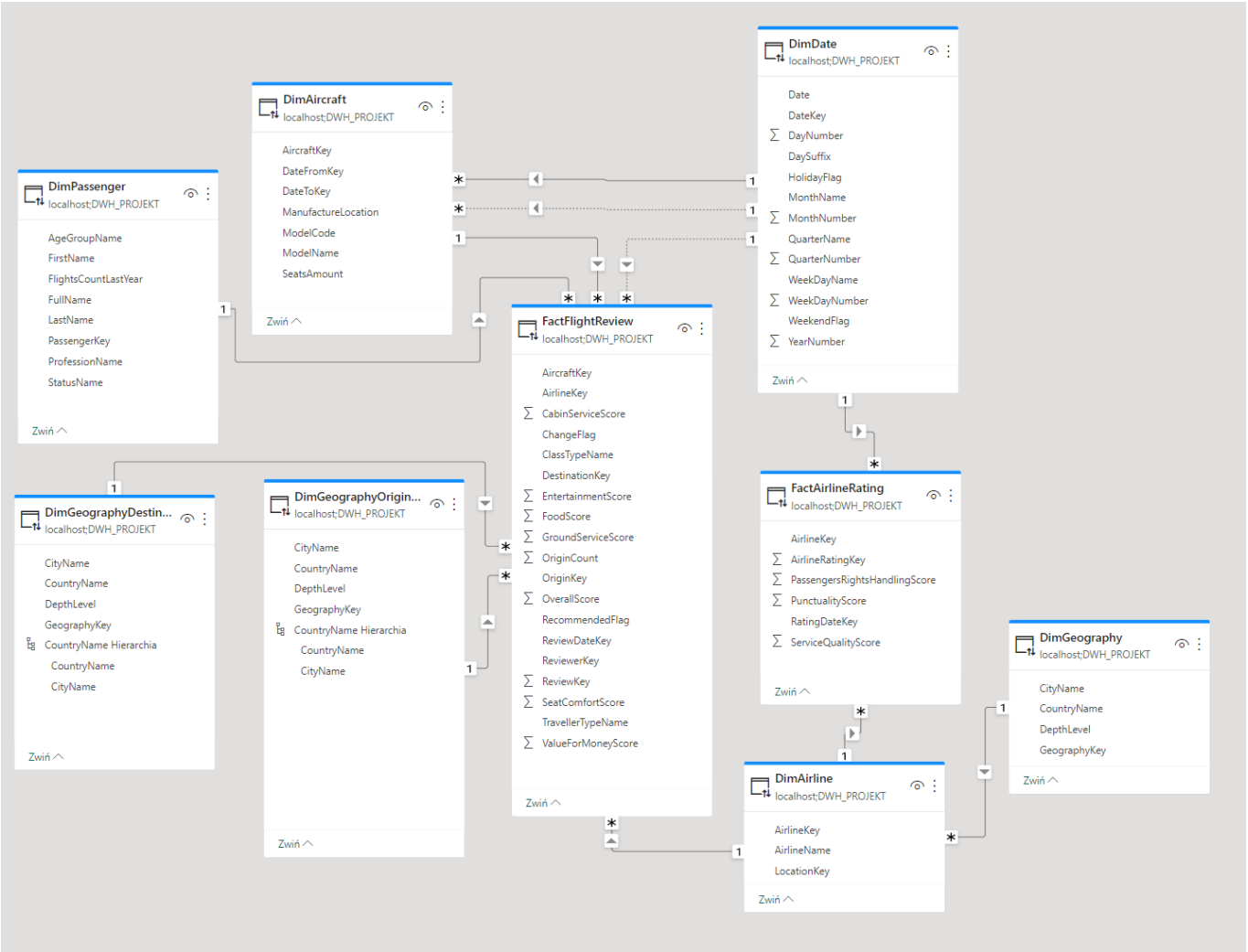
W ramach projektu zostały stworzone 3 raporty opierające się głównie odpowiednio na tabeli faktu FlightReview, tabeli faktu AirlineRating oraz zestawieniu obu tych tabel. Warstwa raportowa została stworzona za pomocą PowerBI. Dostęp do danych w PowerBI opierał się na bezpośrednim połączeniu do hurtowni danych umieszczonej na hoście w trybie DirectQuery.

5.1 Przekształcenia danych w PowerBI

1. Raport pierwszy:
   - zdublowanie wymiaru geografii, aby móc zachować relacje między nim, a faktem z opiniami o lotach, stworzono w ten sposób wymiar geografii dla miasta, z którego lot startował oraz identyczny wymiar dla miasta docelowego
   - stworzono hierarchię dla krajów oraz miast w obydwu wymiarach geograficznych
2. Raport drugi:
   - dodano kolumnę MaxVal w fakcie opinii o liniach lotniczych, która zawsze wynosi 10, aby móc ustawić taką wartość maksymalną w licznikach, które pojawiają się w raporcie
3. Raport trzeci:
   - stworzono kopię wymiaru geografii analogicznie jak w przypadku raportu pierwszego. Dodano także kolejną kopię wymiaru geografii, aby zachować relację między wymiarem linii lotniczej, a geografią
   - dla każdej miarki z faktu opinii o lotach dodano kolumnę z przemnożonymi razy dwa wartościami tej miarki, aby zakres tych miarek dopasował się do zakresu miarek z drugiego faktu tak, aby dało się je porównywać na raportach. Nazwy tych kolumn są takie jak nazwy oryginalnych miarek, ale z przedrostkiem "Scaled"

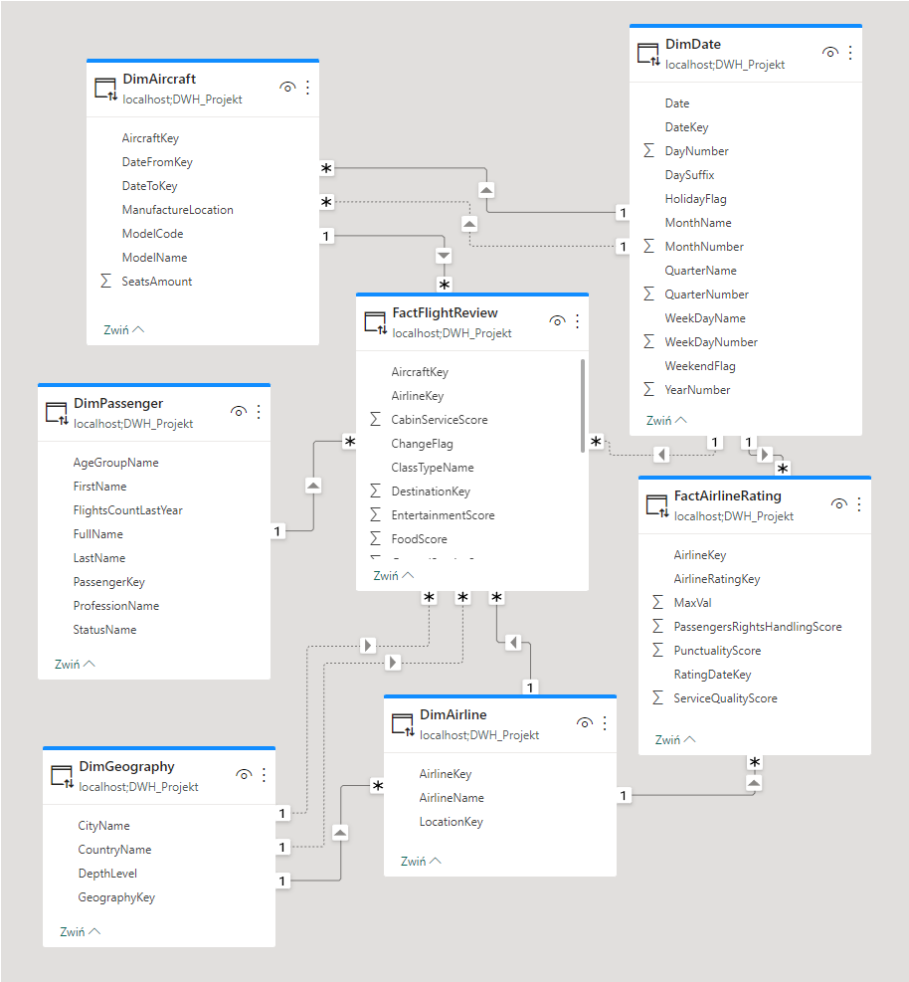
5.2 Modele danych w Power BI

5.2.1 Raport pierwszy



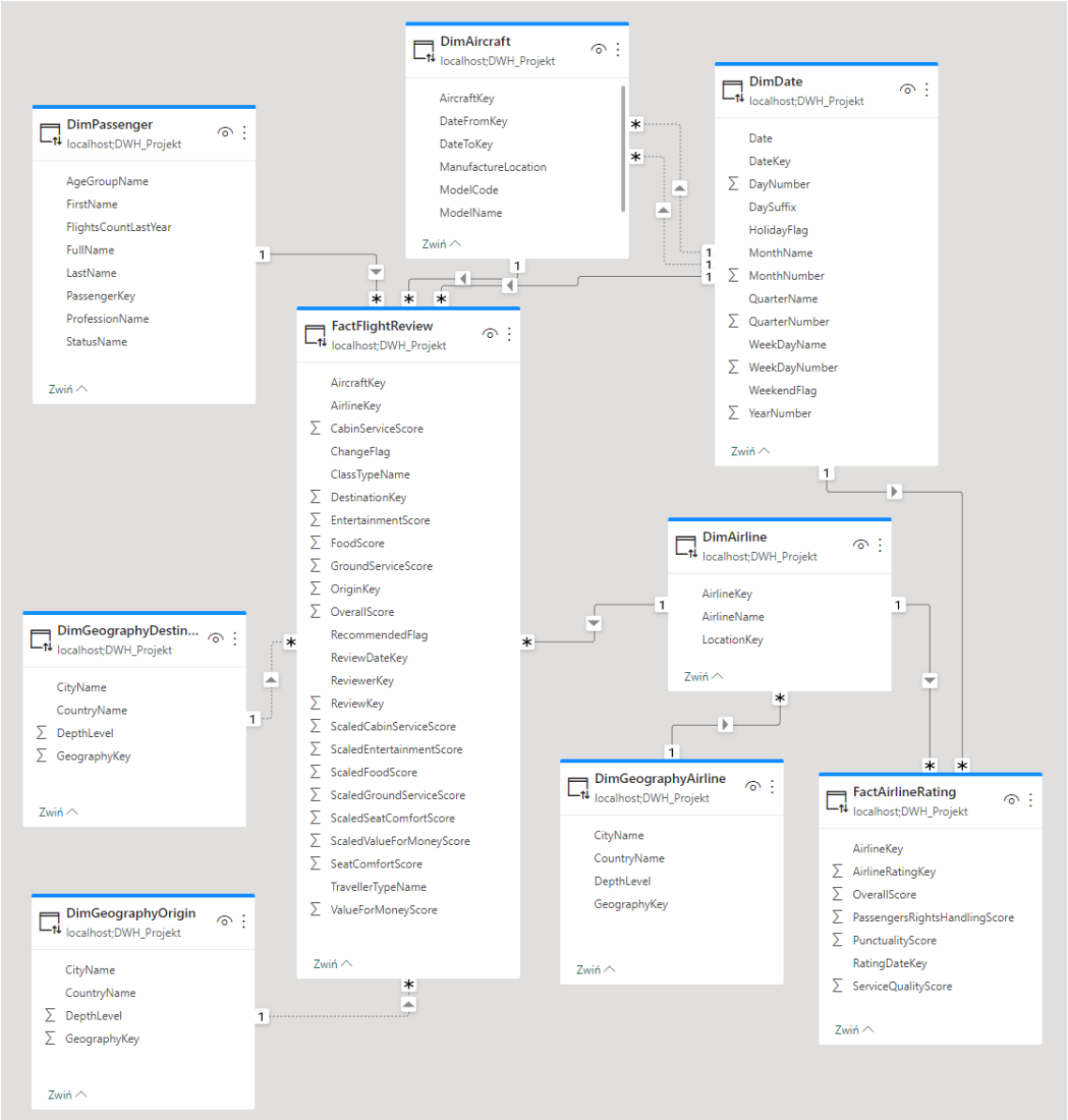
Rysunek 5.1: Diagram modelu danych dla raportu pierwszego

5.2.2 Raport drugi



Rysunek 5.2: Diagram modelu danych dla raportu drugiego

5.2.3 Raport trzeci



Rysunek 5.3: Diagram modelu danych dla raportu trzeciego

5.3 Spójność danych z warstwą hurtowni danych

Aby zbadać spójność danych między warstwą hurtowni, a raportową, sprawdziliśmy ilości rekordów oraz kolumn w tabelach hurtowni danych za pomocą odpowiednich zapytań SQL w Microsoft SQL Server Management Studio oraz w PowerBI za pomocą podgladu danych w widoku modelu. Wartości dla wszystkich tabel dla wszystkich raportów były zgodne, co oznacza, że raporty działają na odpowiednich danych.

5.4 Opis raportów

Użytkownik może filtrować rekordy przedstawiane na wykresach po wybranych przez niego danych.

1. **Pierwszy raport** dotyczy analizy tabel związanych z tabelą faktową FlightReviews.

   - Na pierwszej stronie przeprowadzono analizę opinii recenzentów pod kątem informacji o modelach samolotów ze zbioru Aircraft. Za pomocą wykresu słupkowego porównano modele samolotów pod względem ogólnej oceny wystawianej przez recenzentów o lotach oraz na podstawie wykresu mapy porównano również zależności między ogólną oceną o lotach a miejscem głównej siedziby firmy produkującej dany model samolotu, którym odbywał się lot. Ponadto, przy pomocy wykresu punktowego zestawiono średnie wartości ocen o komforcie siedzeń w samolocie z ilością miejsc dostępnych w modelu samolotu. Z tej analizy można zauważyć, że pasażerowie chwalą wygodę w samolotach o liczbie miejsc w przedziale 200-400 oraz że zarówno najlepsze jak i najgorsze oceny komfortu siedzeń dotyczą modeli samolotów o małej liczbie miejsc pasażerskich. Wśród najlepiej ocenianych modeli samolotów znajdują się w większości modele typu Boeing, Airbus oraz Avro. Dodatkowo, większość najlepiej ocenianych samolotów jest produkowana w Europie.
   - Następnie w opisywanej warstwie raportowej przeanalizowano informacje w zakresie pasażerów wystawiających opinie. Za pomocą wykresu słupkowego sprawdzono zależności między wysokością średniej ocen o zapewnionej rozrywce w czasie lotu wśród różnych grup wiekowych pasażerów na temat danej linii lotniczej. Ponadto, na wykresie kołowym przedstawiono rozkład grup wiekowych pasażerów wystawiających opinie. Następnie porównano liczności recenzentów o danym statusie względem ich ostatecznej decyzji o rekomendacji. Z powyżej opisanej części raportowej można wywnioskować, że osoby starsze mają tendencję do przyznawania wyższych ocen o lotach. Można również zauważyć, że recenzenci to głównie osoby dorosłe w wieku od 25 do 44 lat. Idealny rozkład osób o różnym statusie wynika z faktu, że dane były generowane sztucznie. W rzeczywistym przypadku można by było wyciągać wnioski, czy pasażerom z dużym doświadczeniem z podróżami lotniczymi są rekomendowane loty, które później mogą oni polecić innym.
   - Na ostatniej stronie pierwszego raportu przedstawiono dwie mapy drzewa. Pierwsza obrazuje wysokość średnich ocen o serwisie na lotnisku względem państw, z którego recenzenci wylatywali, a druga pokazuje wysokość ocen względem kraju przylotu. Z tych wykresów można wywnioskować, że pasażerowie najlepiej oceniają obsługę zapewnioną na lotniskach w Azji.

2. **Drugi raport** został stworzony na podstawie tabeli faktowej FactAirlineRanking.

   - Na pierwszej stronie przeprowadzono analizę średniej każdej oceny podanej w rankingu dla wszystkich linii lotniczych na przestrzeni lat 2013-2019 za pomocą wykresu liniowego. Następnie porównano średnie wyniki w zakresie ocen o punktualności oraz ocen o szybkości weryfikacji reklamacji i przetwarzania rekompensat dla każdej linii lotniczej na podstawie wykresu punktowy oraz analogicznego wykresu z tym, że tym razem mamy pogrupowanie na kraj pochodzenia linii lotniczych. W podobny sposób za pomocą wykresu mapy przedstawiono również zestawienie średniej oceny jakości obsługi pasażerów przez każdą linię lotniczą. Z opisanych wykresów można wywnioskować m.in. po 2018 nastąpił spadek w ocenach o punktualności i weryfikacji reklamacji pasażerów średnio dla wszystkich linii lotniczych podczas gdy wysokości ocen o jakości serwisu w czasie lotu wzrosły. Można również zauważyć, że linie lotnicze, które pochodzą z Europy i Azji, są średnio lepiej oceniane w zakresie obsługi praw pasażera oraz jakości obsługi w trakcie podróży niż linie lotnicze wywodzące się z Ameryki Północnej czy Australii.
   - Druga strona raportu dedykowana jest najbardziej aktualnym danym, do jakich mamy dostęp. W tym przypadku jest to rok 2019. Na wykresie słupkowym dane jest porównanie trzech dostępnych miarek, to jest punktualności, jakości serwisu oraz uwzględniania praw pasażera w zależności od linii lotniczej. Miarki tak jak na innych wykresach agregowane są za pomocą średniej, ale tym razem jedynie z roku 2019. Poniżej umieszczono trzy tabelki, które wyświetlają najlepsze linie lotnicze pod kątem poszczególnych miarek oraz wyniki przez nie uzyskane. Na podstawie tych wykresów można przede wszystkim zauważyć, że rankingi jakości serwisu oraz uwzględnienia praw pasażera w 2019 roku są zdominowane przez trzy linie lotnicze: American Airlines, Scandinavian Airlines i Westjet, natomiast w ostatniej kategorii liderzy są całkowicie

inni. Na podstawie wykresu słupkowego widać, że najlepszą linią lotniczą w roku 2019 na podstawie sumy wszystkich wyników jest American Airlines

3. **Trzeci raport** ma za zadanie pokazać relacje między dwoma tabelami faktów w hurtowni

   - Na pierwszej stronie widnieją trzy wykresy, pierwszy z nich to wykres liniowy pokazujący relację między średnim wynikiem ogólnym z opinii pasażerów, a średnim wynikiem ogólnym z opinii o liniach lotniczych. Jak widać po wizualizacji pasażerowie statystycznie dają dość niskie oceny, a tendencja ta z roku na rok jest coraz gorsza. Drugim wykresem jest wykres punktowy przedstawiający podobną zależność jak poprzedni wykres z tym, że tym razem mamy podział na poszczególne linie lotnicze (każda kropka to inna linia lotnicza). Możemy stąd wysnuć identyczne wnioski jak na poprzedniej wizualizacji, ale dodatkowo możemy sprawdzić, czy na przykład oceny na rankingach dla danej linii nie są zawyżone, bądź zaniżone, zestawiając to z opiniami pasażerów. Warto zaznaczyć, że metryki w obydwu faktach są inne, ale mimo to mówią o podobnych aspektach, więc jest sens zestawiać je ze sobą w ten sposób. Ostatni z wykresów przedstawia tę samą zależność co poprzednie dwa, ale tym razem mamy ją od krajów, z których pochodzą linie lotnicze
   - Na kolejnej stronie mamy porównanie średniej ze wszystkich linii lotniczych z wyników jakości serwisu w przeciągu kolejnych lat. Wnioski są analogiczne do tych z poprzedniej strony, ludzie oceniają to znacząco niżej niż rankingi. Dodatkowo mamy tu również porównanie jakości jedzenia oraz rozrywki na pokładzie samolotu przedstawione za pomocą wykresu słupkowego połączonego z liniowym. Jak widać jakość jednej z tych rzeczy idzie w parze z drugą . Kolejną wizualizacją na tej stronie jest wykres punktowy, z którego widzimy relację między średnią dla linii lotniczych punktualnością linii lotniczych, a tym jak bardzo według pasażera warto lot był warty swojej ceny. Ostatnią wizualizacją jest mapa, na której do krajów, z których pochodzą linie lotnicze przyporządkowano ich średni wynik serwisu na lotnisku w postaci rozmiaru bąbelka oraz średnią tego wyniku oddzielnie dla lotów polecanych i nie polecanych przedstawioną za pomocą wykresu kołowego na bąbelkach. Można z niego wywnioskować, że loty, które były polecane przez ludzi miały znacząco lepszy serwis naziemny niż te, które polecane nie były

Rozdział 6

Podsumowanie biznesowych rezultatów projektu

Podsumowując powyższe raporty możemy stwierdzić, że linie lotnicze korzystające z naszej hurtowni byłyby w stanie między innymi mierzyć zależności między poszczególnymi aspektami podróży, a ogólnym zadowoleniem z lotów, bądź prowadzić porównanie między swoimi wynikami, a wynikami konkurencji, co mogłoby ułatwić im rozeznanie w tym jak poprawić jakość świadczonych przez nie usług i dzięki temu zachęcić większą ilość klientów do korzystania z nich. Dodatkowo istnieje możliwość sprawdzenia jaki odbiór wśród klientów mają poszczególne modele samolotów, a więc dzięki temu linie lotnicze mogłyby dowiedzieć się w jakie modele warto zainwestować.

Rozdział 7

Przeprowadzone testy

7.1 Ogólne testowanie

Poza testami wymienionymi poniżej było również testowane ładowanie danych. Każda z tabeli była wielokrotnie usuwana i ładowana ponownie. Przy każdym ładowaniu liczba insertowanych wierszy zgadzała się z ilością rekordów z źródła danych. Sprawdzono również czy przy ponownym ładowaniu bez usuwania tabeli liczba dodanych wierszy jest równa 0. Ponadto, przeprowadzono testy działania transformacji Update przy zmianie rekordów w załadowanych danych źródłowych.

7.2 Data Accuracy

7.2.1 DimAirline

Dla trzech wybranych linii lotniczych sprawdzimy, czy ich kraj pochodzenia zgadza się z faktycznym. Poniższa tabela przedstawia dane wzięte z hurtowni Natomiast faktyczne wartości na podstawie informacji z Wikipedii to:

	AirlineName	CountryName
1	Qatar Airways	Qatar
2	Avianca	Colombia
3	Azul Airlines	Brazil

Rysunek 7.1: Wyniki zapytania testowego znajdującego kraj pochodzenia wybranych liniii lotniczych

- Qatar Airways: Katar
- Avianca: Kolumbia
- AzulAirlines: Brazylia

Tak jak widać dane z hurtowni zgadzają się z rzeczywistymi

7.2.2 DimAircraft

Dla trzech wylosowanych rekordów z DimAircraft zostało sprawdzone czy liczba siedzeń w modelu samolotu i kraj produkcji samolotu zgadza się z faktycznymi wartościami. Prawdziwe informacje o samolotach znalezione w Internecie:

ModelName	ManufactureLocation	SeatsAmount
Shorts SD.360	United Kingdom	36
Boeing 747 all pax models	United States	416
Airbus A310-200 pax	European consortium	198

Rysunek 7.2: Wyniki zapytania testowego sprawdzającego 3 wybrane rekordy z DimAircraft

- Shorts SD.360 - kraj produkcji: UK, liczba siedzeń waha się między 33-36;
- Boeing 747 all pax models - produkowany jest w USA, liczba siedzeń jest estymowana na około 400;

- Airbus A310-200 pax - jest opisany jako projekt międzynarodowy i może pomieścić od 190 do 220 osób.

Stąd widać, że dane z źródła danych Aircraft są adekwatne do rzeczywistych informacji.

7.2.3 FactFlightReview

Dla trzech wybranych rekordów sprawdzimy czy trasa opisana w kolumnie route ze zbioru źródłowego Flight reviews zgadza się z wartościami w kolumnach DestinationKey, OriginKey i ChangeFlag w tabeli FactFlightReview. W poniższej tabeli widzimy porównanie route z kolumnami z hurtowni stworzonych z jej przekształceń: Tak jak widać

	ReviewKey	ChangeFlag	DestinationCityName	OriginCityName	route
1	71	yes	Bangkok	Amsterdam	Amsterdam to Bangkok via Istanbul
2	1139	no	AMS	IST	IST to AMS
3	1923	yes	Bangkok	London	London to Bangkok via IST

Rysunek 7.3: Wyniki zapytania testowego sprawdzającego 3 wybrane rekordy z FactFlightReview

flaga ChangeFlag jest ustawiona na yes tylko jeśli w route widnieje słowo via, które sugeruje, że był to lot z przesiadką. Nazwy miast początkowych oraz końcowych również się zgadzają

7.3 Completeness

7.3.1 Braki danych w FlightReviews

Za pomocą zapytania sprawdzono, czy na podstawie opinii w wynikowej tabeli Aircraft nie pozostały modele samolotów z nazwa modelu jako pusty string (których było dużo w tabeli źródłowej). Zapytanie zwróciło 0 takich wierszy, zatem wszystkie braki danych w nazwach modeli zostały zamienione na Unknown. Ze względu na dużą liczbę braków danych w kolumnach z nazwą linii lotniczej przeprowadzono podobne testy również na tej kolumnie i wymiarze DimAirline. Ten test również potwierdził, że puste stringi zostały zamienione na Unknown.

7.3.2 Test przekształceń danych w wymiarze DimPassenger

W trakcie tworzenia wymiaru stworzono kolumnę Fullname na podstawie kolumn FirstName i LastName. Za pomocą zapytania przeprowadzono test czy przekształcenie nie ucięło danych i kolumna FullName jest dobrą konkatenacją. Wynik testu potwierdził, że przekształcenie zostało przeprowadzone odpowiednio.

7.4 Consistency

7.4.1 Porównanie liczby wierszy w poszczególnych tabelach hurtowni ze źródłem

Stworzyliśmy zapytanie, które na podstawie tabeli źródłowych oblicza liczbę wierszy, które powinny znaleźć się w hurtowni, następnie zestawia je z liczbą wierszy, które znalazły się w wymiarach i faktach po załadowaniu danych. Zestawienie dotyczy tabel: DimAircraft, DimPassenger, DimAirline, FactFlightReview oraz FactAirlineRating.

	TableName	Data	NumberOfRows
1	Aircraft	Source	218
2	DimAircraft	Target	218
3	Passenger	Source	19724
4	DimPassenger	Target	19724
5	Airlines	Source	117
6	DimAirline	Target	117
7	Opinions	Source	64017
8	FactFlightReview	Target	64017
9	Ranking	Source	404
10	FactAirlineRating	Target	404

Rysunek 7.4: Wyniki zapytania testującego liczbę wierszy w źródłach danych i tabelach wynikowych.

7.4.2 Porównanie liczby wierszy po złączeniu tabel faktowych z wymiarami

Przeprowadzono test w formie zapytania SQL polegający na sprawdzeniu, czy po wykonaniu operacji join między tabelą faktową a jej wymiarem liczba wierszy się nie zmienia. Poniżej przedstawiono wyniki testu na Rysunku 7.5.

	FactTable	DimensionTable	Type	NumberOfRows
1	FlightReview	Airline	InFact	64017
2	FlightReview	Airline	AfterJoin	64017
3	FlightReview	Aircraft	InFact	64017
4	FlightReview	Aircraft	AfterJoin	64017
5	FlightReview	Passenger	InFact	64017
6	FlightReview	Passenger	AfterJoin	64017
7	FlightReview	Geography/Origin	InFact	64017
8	FlightReview	Grography/Origin	AfterJoin	64017
9	FlightReview	Geography/Destination	InFact	64017
10	FlightReview	Grography/Destination	AfterJoin	64017
11	FlightReview	Date	InFact	64017
12	FlightReview	Date	AfterJoin	64017
13	AirlineRating	Airline	InFact	404
14	AirlineRating	Airline	AfterJoin	404
15	AirlineRating	Date	InFact	404
16	AirlineRating	Date	AfterJoin	404

Rysunek 7.5: Wyniki zapytania testującego liczbę wierszy po złączeniu tabel faktowych i ich wymiarów

7.5 Uniqueness

7.5.1 Sprawdzenie wystąpień duplikatów wśród unikalnych kolumn

W modelu naszej hurtowni zakładaliśmy unikalność niektórych kolumn w poszczególnych tabelach. W tabeli DimAircraft unikalne powinny być nazwy modeli samolotów (kolumna ModelName), w tabeli DimAirline unikalne miały być nazwy linii lotniczych (kolumna AirlineName), natomiast w wymiarze DimPassenger unikalne jest zestawienie imienia i nazwiska pasażera (kolumna FullName). Stworzyliśmy zapytanie sprawdzające liczbę wierszy w wyżej wymienionych tabelach i zestawiające je z liczbą unikalnych wartości odpowiednich kolumn, których dotyczyło założenie o unikalności.

	TableName	UniqueValues	NumberOfRows
1	DimAircraft	218	218
2	DimAirline	117	117
3	DimPassenger	19724	19724

Rysunek 7.6: Wyniki zapytania testowego sprawdzającego wystąpienia duplikatów

7.6 Validity

7.6.1 Zakres miarek dla faktu FlightReview

Z danych źródłowych dla faktu FlighReview wynika, że miarka OverallScore powinna być z zakresu 1-10, natomiast pozostałe (SeatComfortScore, CabinServiceScore, FoodScore, EntertainementScore, GroundServiceScore i ValueForMoneyScore) powinny być z zakresu 0-5. Poniżej wstawaimy wyniki dla zapytania wyszukującego wartości minimalnych i maksymalnych dla każdej z miarek:

	TableName	Aggregation	OverallScore	SeatComfortScore	CabinServiceScore	MinFoodScore	EntertainmentScore	GroundServiceScore	ValueForMoneyScore
1	FactFlightReview	Min	1	0	0	0	0	0	0
2	FactFlightReview	Max	10	5	5	5	5	5	5

Rysunek 7.7: Wyniki zapytania testowego sprawdzającego zakres miarek dla FactFlightReview

7.6.2 Zakres miarek dla faktu AirlineRating

Z danych źródłowych dla faktu AirlineRating wynika, że miarki PunctualityScore, ServiceQualityScore i PassengersRightsHandlingScore powinny być z zakresu 0-10. Poniżej wstawaimy wyniki dla zapytania wyszukującego wartości minimalnych i maksymalnych dla każdej z miarek:

	TableName	Aggregation	PunctualityScore	ServiceQualityScore	PassengersRightsHandlingScore
1	FactAirlineRating	Min	4,2	0	0,8
2	FactAirlineRating	Max	9,9	10	10

Rysunek 7.8: Wyniki zapytania testowego sprawdzającego zakres miarek dla FactAirlineRating

7.6.3 Porównanie ze źródłem sumy miarek dla faktu FlightReview

Tabela nr 7.9 została wygenerowana jako wynik zapytania porównujący sumy wszystkich rekordów poszczególnych miarek faktu FlightReview z sumą odpowiadających kolumn z tabeli źródłowej.

	Data	SumOverallScore	SumSeatComfortScore	SumCabinServiceScore	SumFoodScore	SumEntertainmentScore	SumGroundServiceScore	ValueForMoneyScore
1	Target	329395	176124	191340	150841	124895	105980	185983
2	Source	329395	176124	191340	150841	124895	105980	185983

Rysunek 7.9: Wyniki zapytania testowego porównującego ze źródłem sumy miarek dla FactFlightReview

7.6.4 Porównanie ze źródłem sumy miarek dla faktu AirlineRating

Tabela na Rysunku 7.10 została wygenerowana jako wynik zapytania porównujący sumy wszystkich rekordów poszczególnych miarek faktu AirlineRating z sumą odpowiadających kolumn z tabeli źródłowej.

	Data	SumPunctualityScore	SumServiceQualityScore	SumPassengersRightsHandlingScore
1	Source	3180,7	2770,1	2734,6
2	Target	3180,7	2770,1	2734,6

Rysunek 7.10: Wyniki zapytania testowego porównującego ze źródłem sumy miarek dla FactAirlineRating

7.6.5 Testy kolumn kategorycznych

Dla kolumn kategorycznych przeprowadziliśmy test polegający na sprawdzeniu, czy wszystkie kategorie, które oczekiwaliśmy znajdują się w danych. Zrobiliśmy to za pomocą wywołań komendy DISTINCT na poszczególnych kolumnach kategorycznych. Wyniki zgadzały się ze spodziewanymi wartościami.

7.7 Test SCD2

Tabela DimAircraft została stworzona jako SCD2. Aby sprawdzić działanie ETL w jej przypadku, przeprowadziliśmy modyfikację jednego rekordu, po czym ponownie załadowaliśmy dane do tego wymiaru. Wynik dla zmienianego rekordu można zauważyć w tabeli na Rysunku 7.11.

```
update DimAircraft
set DateFromKey = '20230101'
where AircraftKey = 10225;
update DimAircraft
set ModelCode = 'ABC'
where AircraftKey = 10225;
```

	AircraftKey	ModelName	ModelCode	SeatsAmount	ManufactureLocation	DateFromKey	DateToKey
1	10225	Airbus A300 pax	ABC	200	European consortium	20230101	20230610
2	10438	Airbus A300 pax	A30B	200	European consortium	20230610	99991231

Rysunek 7.11: Wyniki zapytania testowego porównującego rekord aktualny z zamkniętym dla zmodyfikowanego samolotu

Rozdział 8

Podział pracy w zespole

Autorami projektu, implementacji rozwiązania, przeprowadzenia testów, stworzenia dokumentacji są Grzegorz Zbrzeżny, Izabela Telejko i Anna Wawrzyńczak.