Hurtownie Danych – Projekt Airbnb

Martyna Majchrzak, Agata Makarewicz

Cel projektu

Stworzenie hurtowni z danych pochodzących z ogłoszeń serwisu Airbnb zawierającej informacje o mieszkaniach wynajmowanych w różnych miastach na całym świecie.

Opis celu biznesowego

Stworzenie zestawienia pozwalającego ocenić potencjalnym wynajmującym jakiej ceny mogą oczekiwać za mieszkanie w ich standardzie, jakie mieszkania są najchętniej wynajmowane oraz najlepiej płatne.

Wykorzystywane zbiory danych

- **Airbnb**

Źródło: http://insideairbnb.com/get-the-data.html

Udostępniane dane zawierają m.in. informacje o lokalizacji, właścicielu, dostępności oraz cenie danego mieszkania/domu, a także oceny użytkowników, którzy je wynajmowali.

Dane pozyskiwane są metodą scrapingu z ogłoszeń na portalu Airbnb i są aktualizowane mniej więcej co miesiąc. Dane dla każdej lokalizacji i dla każdego miesiąca znajdują się w osobnych tabelach. W momencie, gdy pojawiają się dane z nowego miesiąca tabela z poprzedniego zostaje dodana do zakładki ‘archived data’.

Wszystkie wykorzystywane przez nas dane znajdując się w plikach ‘listings.csv’, każdy dla konkretnej lokalizacji.

Pobierzemy dane dla 21 lokalizacji z Hiszpanii, Portugalii i Włoszech.

1. Barcelona, Catalonia, Spain
2. Euskadi, Euskadi, Spain
3. Girona, Catalonia, Spain
4. Madrid, Comunidad de Madrid, Spain
5. Malaga, Andalucía, Spain
6. Mallorca, Islas Baleares, Spain
7. Menorca, Islas Baleares, Spain
8. Sevilla, Andalucía, Spain
9. Valencia, Valencia, Spain
10. Lisbon, Lisbon, Portugal

11. Porto, Norte, Portugal
12. Bergamo, Lombardia, Italy
13. Bologna, Emilia-Romagna, Italy
14. Florence, Toscana, Italy
15. Milan, Lombardy, Italy
16. Naples, Campania, Italy
17. Puglia, Puglia, Italy
18. Rome, Lazio, Italy
19. Sicily, Sicilia, Italy
20. Trentino, Trentino-Alto Adige/Südtirol, Italy
21. Venice, Veneto, Italy

- **Eurostat**

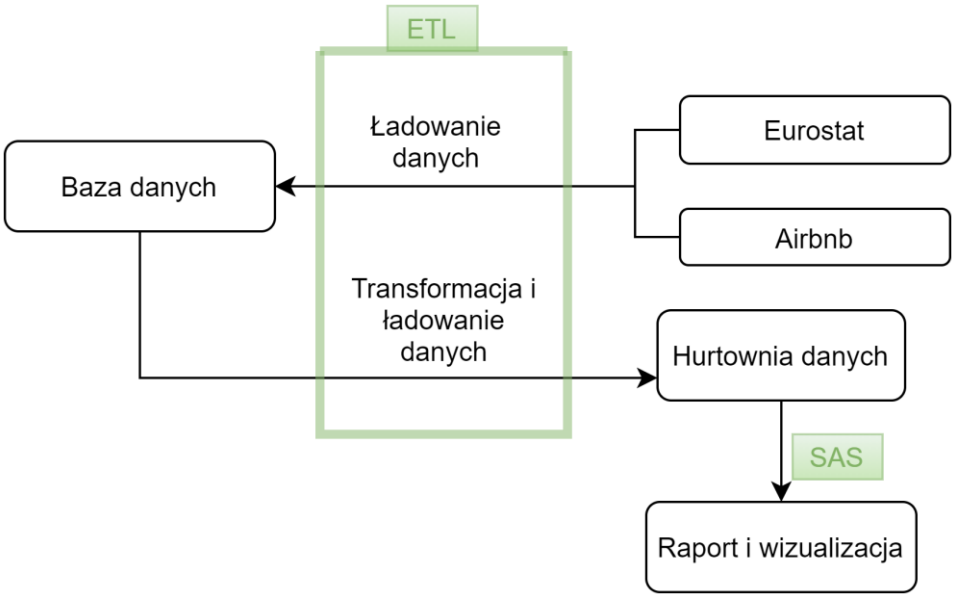
Źródło: https://ec.europa.eu/eurostat/web/regions/data/database

Skorzystamy z danych:

1. Regional Tourism Statistics
    - Arrivals at tourist accommodation establishment, 2019
    - Number of establishments, bedrooms and bed-places, 2019
2. Regional Transport Statistics
    - Air transport of passengers, 2019
3. Reginal Demographic Statistic
    - Population on 1 January, 2020

Dane te zostaną pobrane dla wymienionych w poprzednim punkcie 21 lokalizacji.

Architektura rozwiązania



Rys. 1 Diagram architektury rozwiązania

Dane z poszczególnych tabel z serwisu Airbnb zostaną załadowane do jednej, wspólnej tabeli *Listings* w bazie danych, a dane z serwisu Eurostat do 4 tabel odpowiadających 4 opisanym wcześniej miarom.

Następnie na danych zostaną w odpowiednich procesach ETL dokonane transformacje i tak przygotowane zostaną one załadowane do hurtowni.

W ostatnim kroku stworzymy raport dla użytkownika końcowego z wizualizacjami danych z hurtowni.

Model hurtowni danych

Model hurtowni danych stworzony został zgodnie ze schematem gwiazdy.

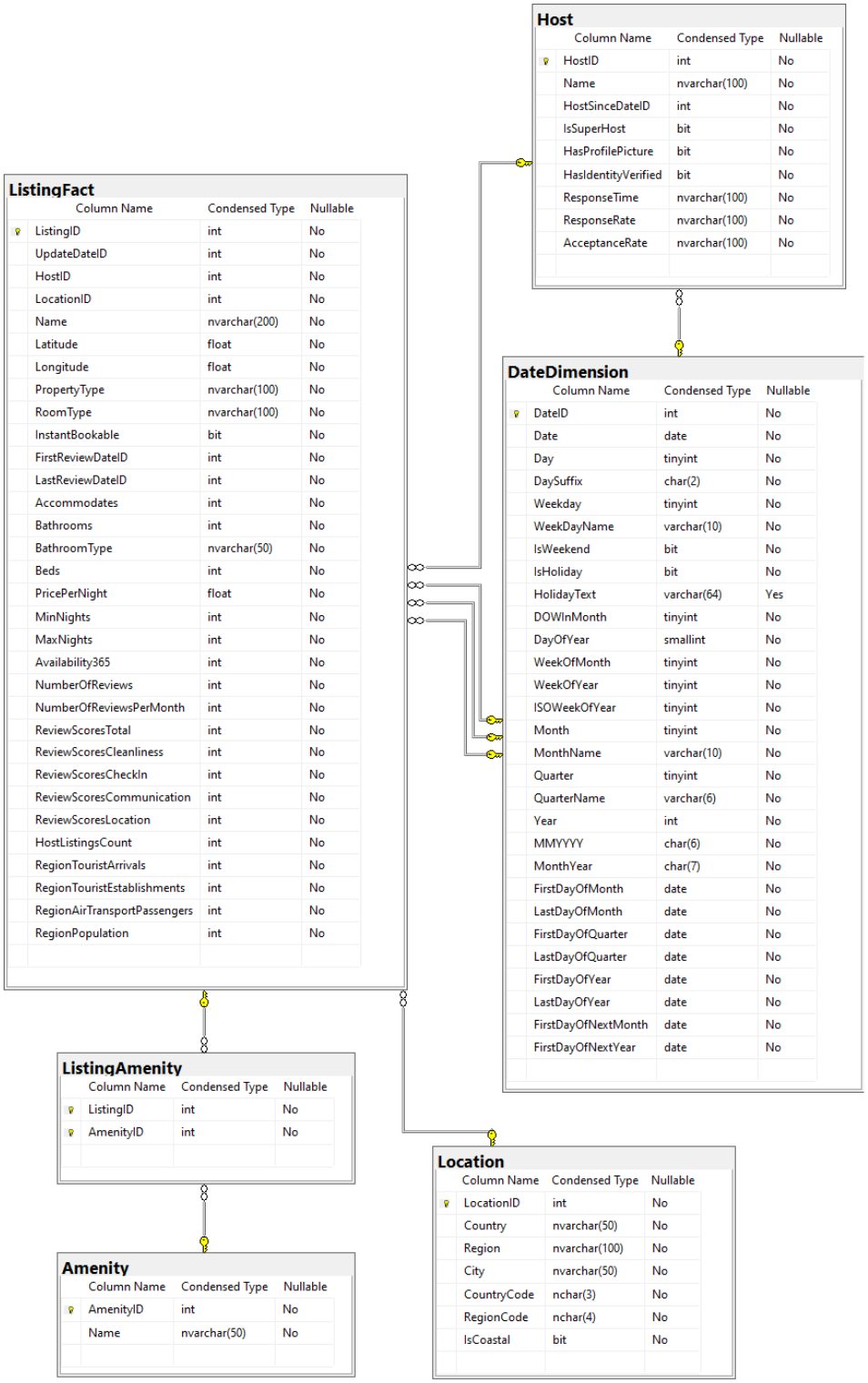
Tabela faktów:

- ListingFact - zawiera pojedyncze ogłoszenia (miejsca pobytu) z informacjami takimi jak data pobrania danych, typ budynku/pomieszczenia, długość i szerokość geograficzna, oraz z kluczami obcymi do opisujących je tabel wymiarów. Oprócz tego tabela faktów zawiera następujące miary:
    - Accomodates – liczba miejsc
    - Bathrooms – liczba łazienek
    - Beds – liczba łóżek
    - PricePerNight – cena za noc
    - MinNights, MaxNights – minimalna, maksymalna liczba noclegów
    - Availability365 - dostępność przez rok od daty pobrania danych
    - NumberOfReviews, NumberOfReviewsPerMonth - liczba recenzji
    - ReviewScores (Total, Cleanliness, CheckIn, Communication, Location) - oceny użytkowników (całkowite i w podziale na kategorie)

- HostListingsCount – liczba ogłoszeń gospodarza
- RegionTouristArrivals - liczba przyjeżdżających turystów
- RegionTouristEstablishments - liczba instytucji turystycznych
- RegionAirTransportPassengers - liczba pasażerów linii lotniczych
- RegionPopulation - populacja

Tabele wymiarów:

- Amenity – opisuje udogodnienia, które oferowane są w danym miejscu pobytu (np. WiFi, kuchnia, parking itp.). Zawiera klucz danego udogodnienia (klucz główny) oraz jego nazwę. W celu obsłużenia relacji wiele-do-wielu (jedno miejsce może oferować wiele udogodnień, a jedno udogodnienie może być dostępne w wielu miejscach) utworzona została tzw. *bridge table* – ListingAmenity - która przechowuje pary postaci: *klucz ogłoszenia (miejsca) - klucz udogodnienia*.
- Host - zawiera informacje o gospodarzu danego miejsca pobytu takie jak imię i nazwisko, datę rejestracji w serwisie oraz flagi – czy posiada status *'SuperHost'*, zdjęcie profilowe i czy zweryfikował tożsamość. Oprócz tego dodane zostały cechy dotyczące czasu i procentu odpowiedzi, oraz akceptacji zgłoszenia chęci wynajmu.
- Location – opisuje lokalizację danego miejsca pobytu poprzez hierarchię *państwo - region – miasto,* wraz z dodatkowymi informacjami: kodem kraju i regionu oraz informacją czy dany region jest regionem nadmorskim.
- DateDimension – opisuje wszelkie daty zawarte w hurtowni, zarówno w tabeli faktów, jak i wymiarze gospodarza (jest to zatem tzw. r*ole-playing dimension*), w których to pola odnoszące się do daty zawierają tylko jej klucz, po którym łączymy się z tym wymiarem. Zawiera informacje uszczegółowiające datę takie jak czy jest to weekend, który to tydzień roku itp. .



Rys. 2 Model hurtowni danych

Transformacje danych w procesach ETL

1. Stworzenie słownika - wymiaru *Amenity*

W bazie danych informacje o udogodnieniach danego miejsca pobytu przechowywane są jako lista wartości w poszczególnych wierszach. Dane te trzeba

rozbić na pojedyncze pola i stworzyć listę unikatowych wartości, które znajdą się w tabeli *Amenity*. Następnie ogłoszenia zostaną połączone z tym wymiarem za pomocą *bridge table*.

2. Konwersja miar dotyczących hosta do kategorycznych

W tabeli Host znajdą się te informacje z tabeli Listing, które dotyczą wynajmującego dane miejsce pobytu. Dwa z nich: *ResponseRate* i *AcceptanceRate* to wartości procentowe, które należy pogrupować w kategorie:

- 0-49% - low
- 50-79% – medium
- 80-99% – high
- 100 %- max

3. Konwersja wartości t/f na bit

Wartości *IsSuperHost*, *HasProfilePicture, HasIdentityVerified* oraz *InstantBookable* w bazie danych kodowane są literami t- true i f- false. Należy zrzutować je na typ bit, gdzie 1 będzie oznaczało true, a 0 false.

4. Konwersja dat

W bazie danych daty przechowywane są w formacie DD.MM.YYYY. Wykorzystamy DateDimension tak, aby w poszczególnych polach w hurtowni przechowywane było DateID z tego wymiatu (int).

5. Utworzenie sztucznego klucza *LocationID*

W tabeli faktów informacje o lokalizacji przechowywane są w polu *Neighbourhood* w formie "Państwo, region, miasto". Zastąpione zostanie ono sztucznym kluczem, po którym będziemy łączyć fakty z wymiarem *LocationID*.

Warstwa raportowa

Wizualizacja danych przygotowana zostanie przy pomocy narzędzia SAS Visual Analytics. Przestawione zostaną między innymi:

- Wizualizacja poszczególnych miejsc pobytu na mapie wraz z informacjami o nich w postaci tooltipów
- Liczba ogłoszeń w danym regionie w zależności od ilości przyjeżdżających do niego turystów/liczby placówek turystycznych
- Ceny wynajmu w zależności od ocen użytkowników
- Ceny wynajmu w zależności od cech gospodarza

Do wizualizacji dodana zostanie możliwość filtrowania po lokalizacji (hierarchicznie), rodzaju budynku/pokoju, ocenach lub też cechach gospodarza.