# Hurtownie danych i systemy Business Inteligence
# Analiza danych zasięgu sieci telefonicznej w Polsce

Karol Degórski i Jakub Lis

17 maja 2022

## 1 Opis biznesowy

Celem projektu jest przeanalizowanie zasięgu sieci komórkowej w Polsce. W dzisiejszych czasach coraz bardziej popularna staje się być sieć nowej generacji 5G. Dzieje się tak mimo licznych kontrowersji, mitów czy fake newsów z nią związanych. Przykładami takich nieprawdziwych wiadomości może być:

- telefon komórkowy pod wpływem sieci 5G działa jak mikrofalówka przez co podgrzewa nasz mózg;
- sieć 5G pozwala na śledzenie nas przez cały czas;
- sieć 5G działa rakotwórczo.

Jednakże znacząca większość społeczeństwa dostrzega ogromne zalety korzystania z sieci nowej generacji takie jak:

- znacząco szybsze połączenie;
- bardziej stabilny dostęp do sieci;
- krótsze czasy reakcji sieci.

Dlatego też celem naszego projektu jest przeanalizowanie jakie grupy ludności Polski znajdują się w zasięgu tej sieci. Dokonamy analizy dotyczącej takich kwestii jak np. gęstość zaludnienia, struktura wiekowa, wykształcenie itd. Wykorzystamy dane o masztach transmisyjnych pochodzące ze strony: http://beta.btsearch.pl. Zawierają one informację o operatorze, wykorzystywanych technologiach oraz lokalizacji danego masztu. Aby uzyskać dane ekonomiczne skorzystamy z bazy danych GUS (Bank Danych Lokalnych). Połączone w ten sposób dane zwizualizujemy na mapie Polski z wykorzystaniem szeregu wykresów do porownań. Analiza wydaje się być kluczowa głównie z punktu widzenia operatorów telefonii komórkowej, ponieważ pozwoli ona na znalezienie nowych grup użytkowników, którzy mogą być zainteresowani uzyskaniem dostępu do tej technologii. Dzięki temu zyski firm telekomunikacyjnych mogą zostać zwiększone. Ponadto dana firma może porównać się do konkurencji i sprawdzić jak jej zasięg wypada w porównaniu z innymi. Zapewnimy również rzetelny dostęp do informacji dla zwykłych ludzi, aby mogli sprawdzić, czy mają dostęp do danej sieci oraz poszerzyć swoją wiedzę o ciekawe statystyki.

# 2 Opis danych

## 2.1 Dane dotyczące masztów sieciowych

Źródłem danych są wyeksportowane pliki CLF ze strony `http://beta.btsearch.pl/bts/` dotyczące masztów sieciowych. W danych znajdują się takie informacje, jak

- *LAT i LON* - współrzędne geograficzne masztu;
- *POS-RAT* - dokładność współrzędnych miejsca, wartością -1 oznaczone są dokładnie podane współrzędne;
- *DESC* - opis w postaci adresu i informacji na temat masztu, np. że jest na dachu budynku, albo że jest osobną wieżą; także informacja nt. operatora masztu (T-mobile, Orange, itd.);
- *SYS* - kod odpowiadający standardowi masztu, np. GSM, LTE, 5G;

Dokładniejszy opis tych zmiennych jest dostępny na stronie `https://sites.google.com/site/clfgmon/clf4?authuser=0`, ale w przypadku niewymienionych kolumn najczęściej dane zawierają wartość -1, oznaczającą brak informacji.

## 2.2 Dane ekonomiczne pochodzące z GUS

Drugim źródłem danych jest Bank Danych Lokalnych GUS. Zawiera on bardzo dużo różnych tabel dotyczących ludności Polski. Dane te są w postaci plików CSV. Wybraliśmy kilka najbardziej interesujących tabel, są to:

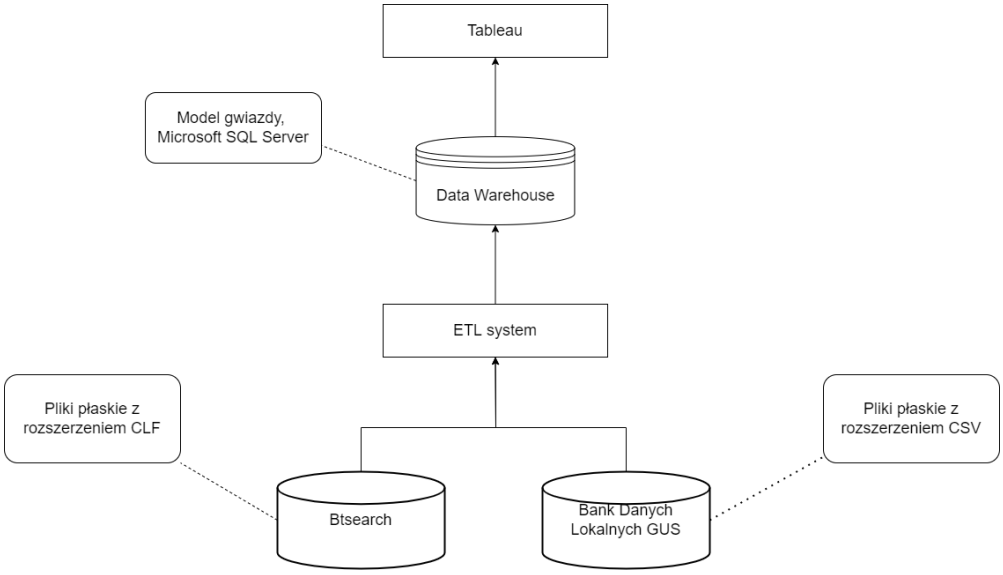
- Gęstość zaludnienia - ludność na $1km^2$ na poziomie gmin;
- Ludność według grup wieku i płci - dane na poziomie gmin;
- Bezrobocie zarejestrowani wg wieku - dane na poziomie powiatów;
- Ludność w wieku 13 lat i więcej wg grup wieku i poziomu wykształcenia - dane na poziomie powiatów;
- Wyposażenie w niektóre przedmioty trwałego użytkowania w % ogółu gospodarstw domowych - dane na poziomie województw;

# 3 Opis architektury

W niniejszym punkcie przedstawimy opis architektury hurtownii danych

## 3.1 Podział na warstwy

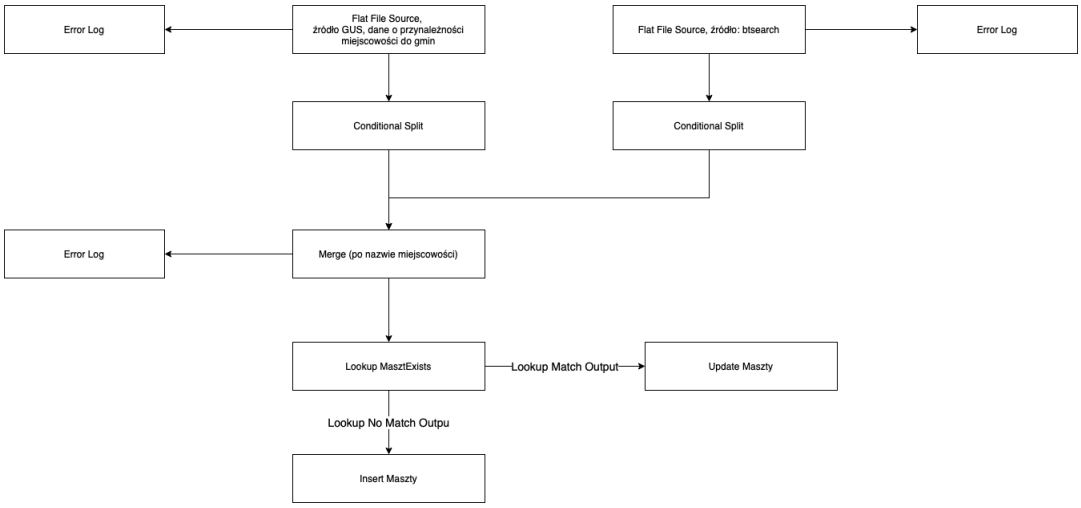
Architekrurę DWH/BI prezentujemy na rysunku 1. Najniższą warstwę stanowią dwie bazy danych: btsearch (pliki CLF) oraz Bank Danych Lokalnych GUS (pliki csv). Następnie dane są łączone oraz obrabiane poprzez proces ETL, który omówimy w kolejnym punkcie. Dane powstałe w ten sposób zasilają hurtownię danych w modelu gwiazdy w Microsoft SQL Server. Ostatnią warstwą jest warstwa narzędzia klasy Business Inteligence, w naszym przypadku Tableau.




Rysunek 1: Architektura DWH/BI

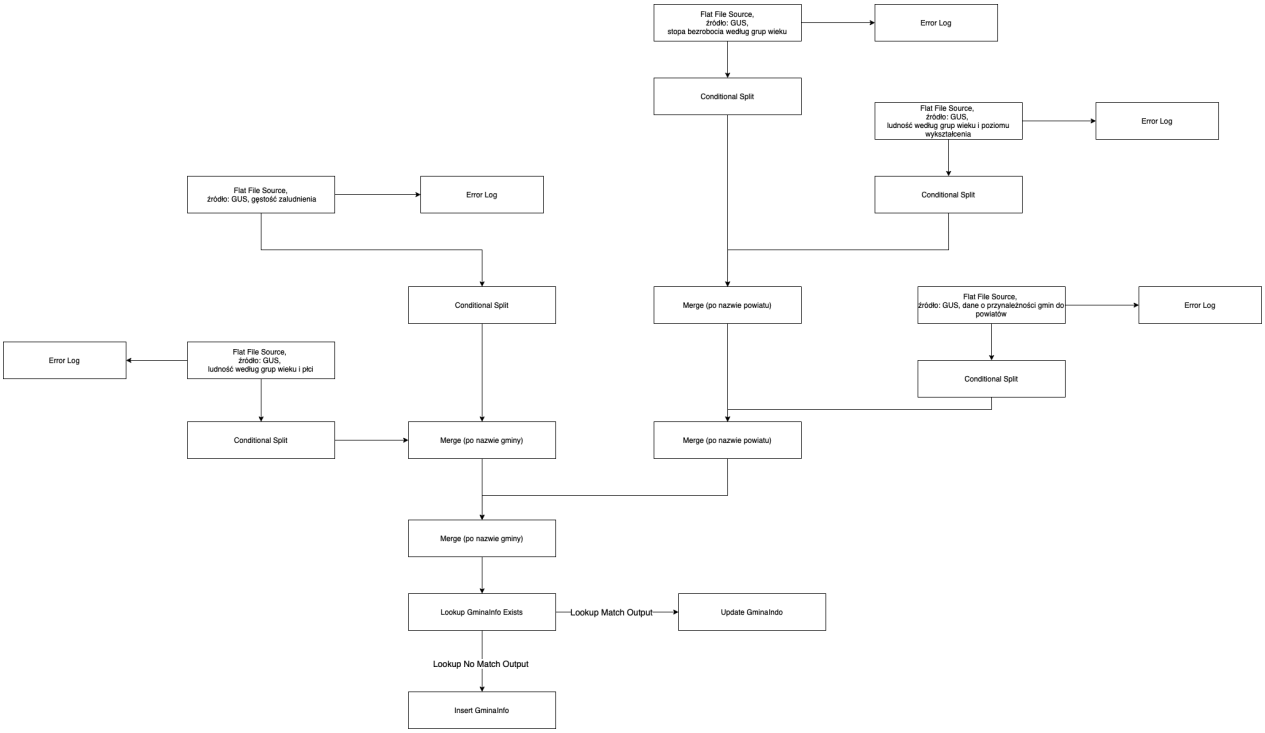
## 3.2 Proces ETL

W tym punkcie omówimy jak wygląda proces ETL, który dostarcza dane do naszej hurtowni. Proces dla danych pochodzących z bazy danych o masztach prezentujemy na rysunku 2. Pierwszym krokiem jest wczytanie danych z pliku CLF. Następnie dokonujemy podziału tego pliku na poszczególne kolumny. W przypadku błędnych rekordów zapisujemy taką informację do ErrorLog. Następnie dokonujemy dodania nazwy gminy. Na koniec sprawdzamy czy dany rekord o danym maszcie jest już w bazie danych: jeśli jest to go aktualizujemy, a jeśli nie to dodajemy nowy. W ten sposób zasilana jest tabela Maszty.




Rysunek 2: Schemat procesu ETL dla danych o masztach

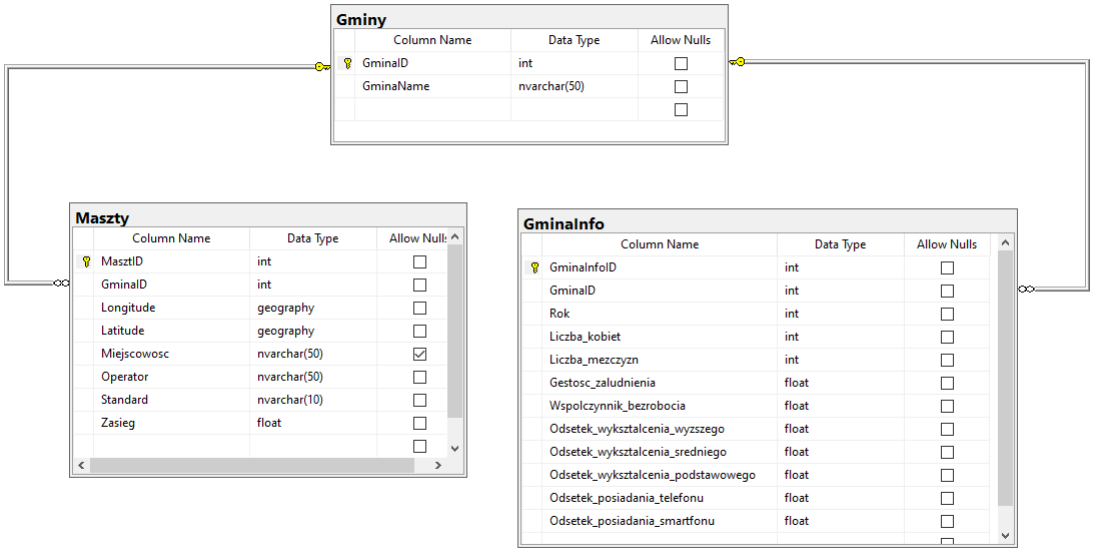
Proces ETL dla danych pochodzących z Banku Danych Lokalnych GUS prezentujemy na rysunku 3. Dane te również rozdzielamy na kolumny. Następnie dokonujemy merge danych pochodzących z różnych plików csv, ale dostępnych na tym samym poziomie jednostki administracyjnej. Te kroki wykonujemy zarówno dla danych dostępnych na poziomie gmin jak i powiatów. Dla danych dostępnych jedynie na poziomie powiatów dokonujemy połączenia z informacją, które gminy znajdują się w danym powiecie. Dzięki temu uzyskujemy estymację tych wartości na poszczególne gminy. Następnie dokonujemy połączenia tych tabel z tabelami, które pierwotnie zawierały informacje o gminach. Analogicznie, jak w przypadku tabeli Maszty sprawdzamy, czy dany rekord jest już naszej bazie. Jeśli jest to go aktualizujemy, a jeśli nie to dodajemy nowy. W ten sposób uzyskujemy dane do tabeli GminaInfo.




Rysunek 3: Schemat procesu ETL dla danych o gminach

## 3.3 Model hurtowni danych

Na rysunku 4 przedstawiamy model hurtownii danych. Hurtownia ta jest zbudowana w modelu gwiazdy. Tabelą faktów w naszej bazie danych jest tabela Gminy. Ponadto znajdują się w niej dwie tabele wymiarowe: Maszty i GminaInfo. Tabela Maszty zawiera informacje o masztach telefonicznych takie jak ich współrzędne geograficzne, standard, miasto, zasięg, czy nazwa operatora. Z kolei tabela GminaInfo zawiera podstawowe informacje o mieszkańcach danej gminy. Są to między innymi liczba kobiet i mężczyzn, gęstość zaludnienia, współóczynnik bezrobocia, informacje o wykształceniu, czy informacje o posiadaniu telefonów przez mieszkańców.




Rysunek 4: Model hurtowni danych