ZADANIE PROJEKTOWE Z HURTOWNI DANYCH - 2020

Dominik DAWIDZIAK, Patryk BARCZAK
gr. I7B1S1

1. Opis założeń biznesowych

Hurtownia danych ma umożliwiać analizę tempa rozprzestrzeniania się wirusa COVID-19 na świecie.

1.1. Wymiary analizy

- Geografia - (kontynent, kraj, populacja, PKB),
- Czas - (rok, miesiąc, dzień),
- Czas od pierwszej detekcji wirusa w danym kraju (numer kolejny dnia),
- Pacjent - (wiek, płeć) - dane generowane losowo

1.2. Miary

- liczba zakażeń (w okresie),
- liczba zgonów (w okresie),
- liczba pacjentów wyleczonych (w okresie),
- liczba nowych przypadków (w danym dniu),
- liczba zakażonych (stan na dzień),
- dynamika zakażeń (liczba nowych przypadków w dniu dzisiejszym/liczba nowych przypadków w dniu wczorajszym)

1.3. Założenia

- Za datę rozpoczęcia rozprzestrzeniania się wirusa, przyjmuje się 21.01.2020 r.
- W przypadku błędów w danych (a takie występują) i pojawianiem się ujemnych wartości np. nowych przypadków zachorowań w danym dniu - co niesie za sobą też uzyskanie ujemnej dynamiki. Wszystkie wartości ujemne mają zostać zastąpione wartością 0.
- Proces ETL ma łączyć się online do źródeł danych,
- Czas pierwszej infekcji wirusem mierzony indywidualnie dla kraju,
- Dane o pacjentach mają być generowane losowo,
- Pacjent który zachorował w kraju X nie może zostać wyleczony ani umrzeć w kraju Y.

2. Dokumentacja

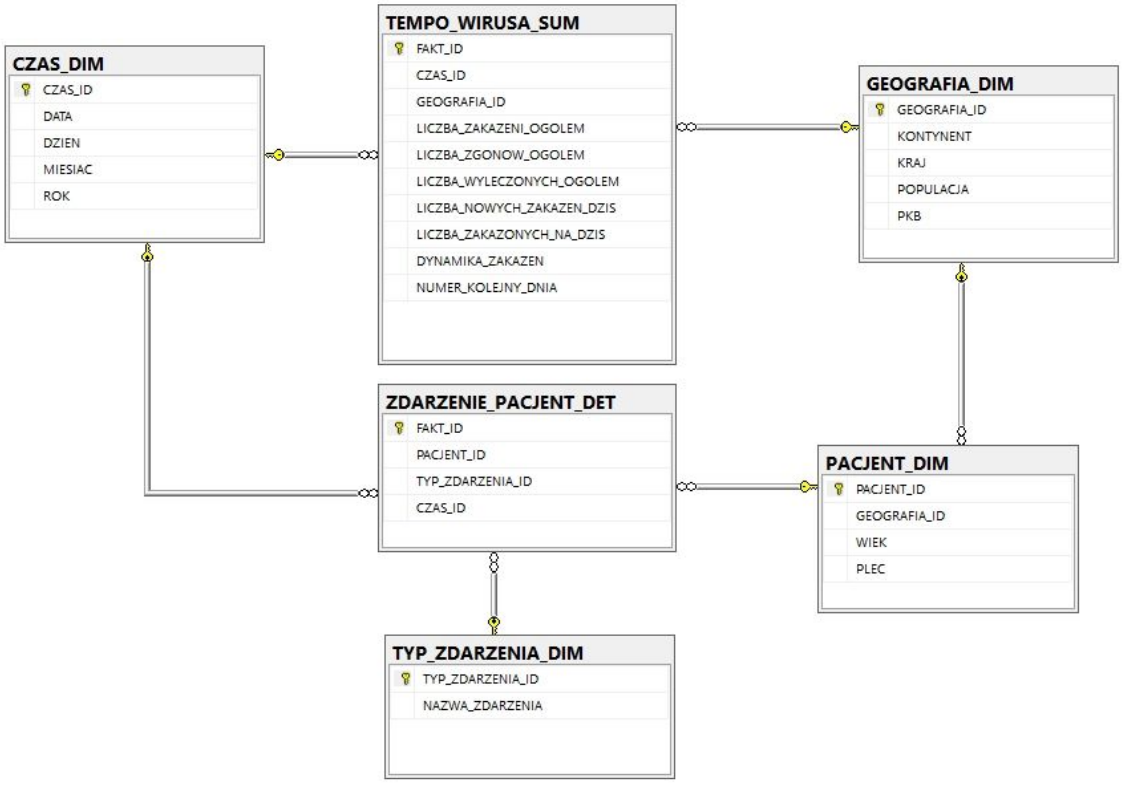
2.1. Dane źródłowe

Źródło zapewniające dane w zakresie zagregowanej liczby przypadków
https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series

Liczba zakażonych (gradacja dzienna)	time_series_covid19_confirmed_global.csv
Liczba wyleczonych (gradacja dzienna)	time_series_covid19_recovered_global.csv
Liczba zgonów (gradacja dzienna)	time_series_covid19_deaths_global.csv

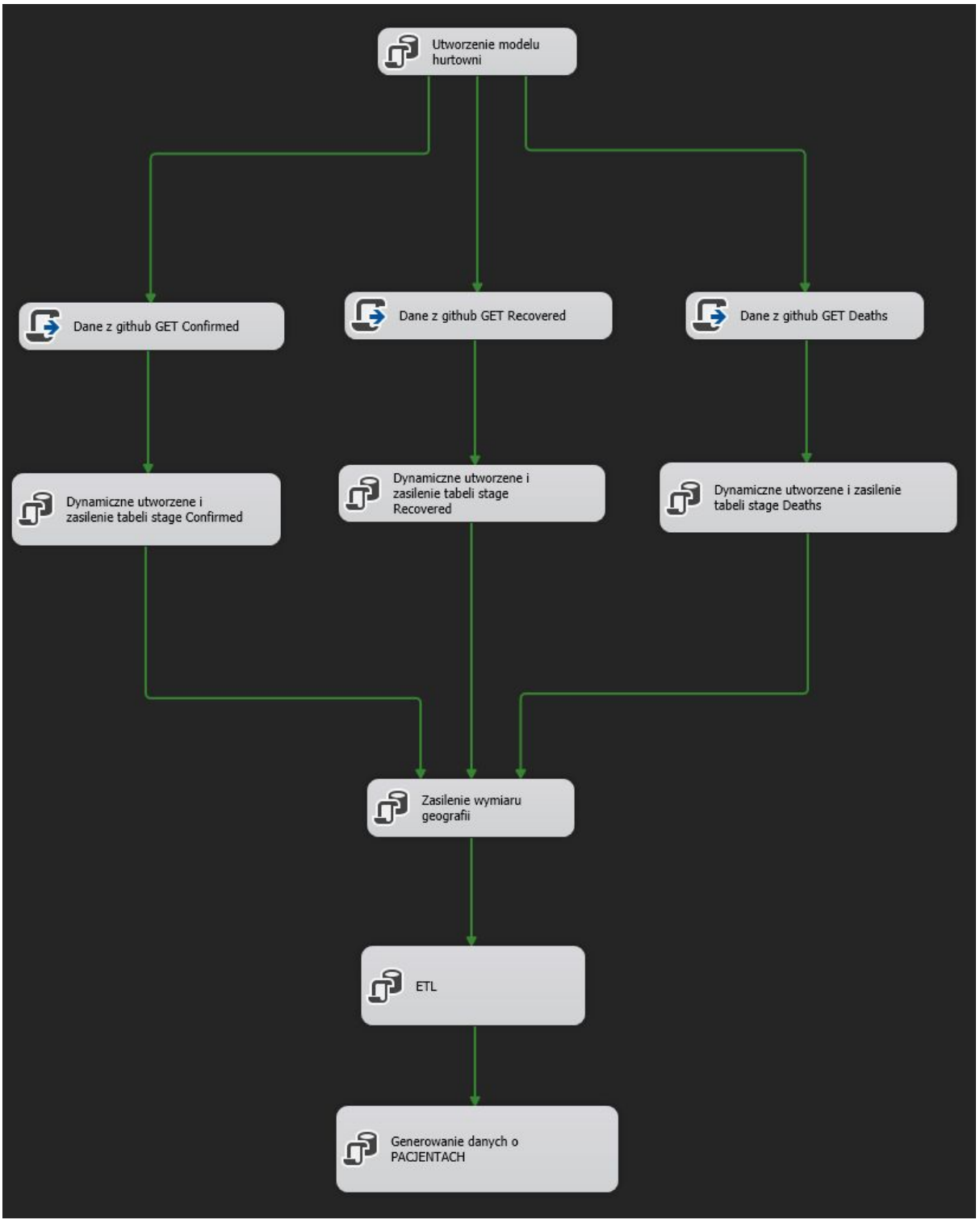
Dane dotyczące pacjentów są generowane losowo na podstawie liczby przypadków

2.2. Model wymiarowy hurtowni

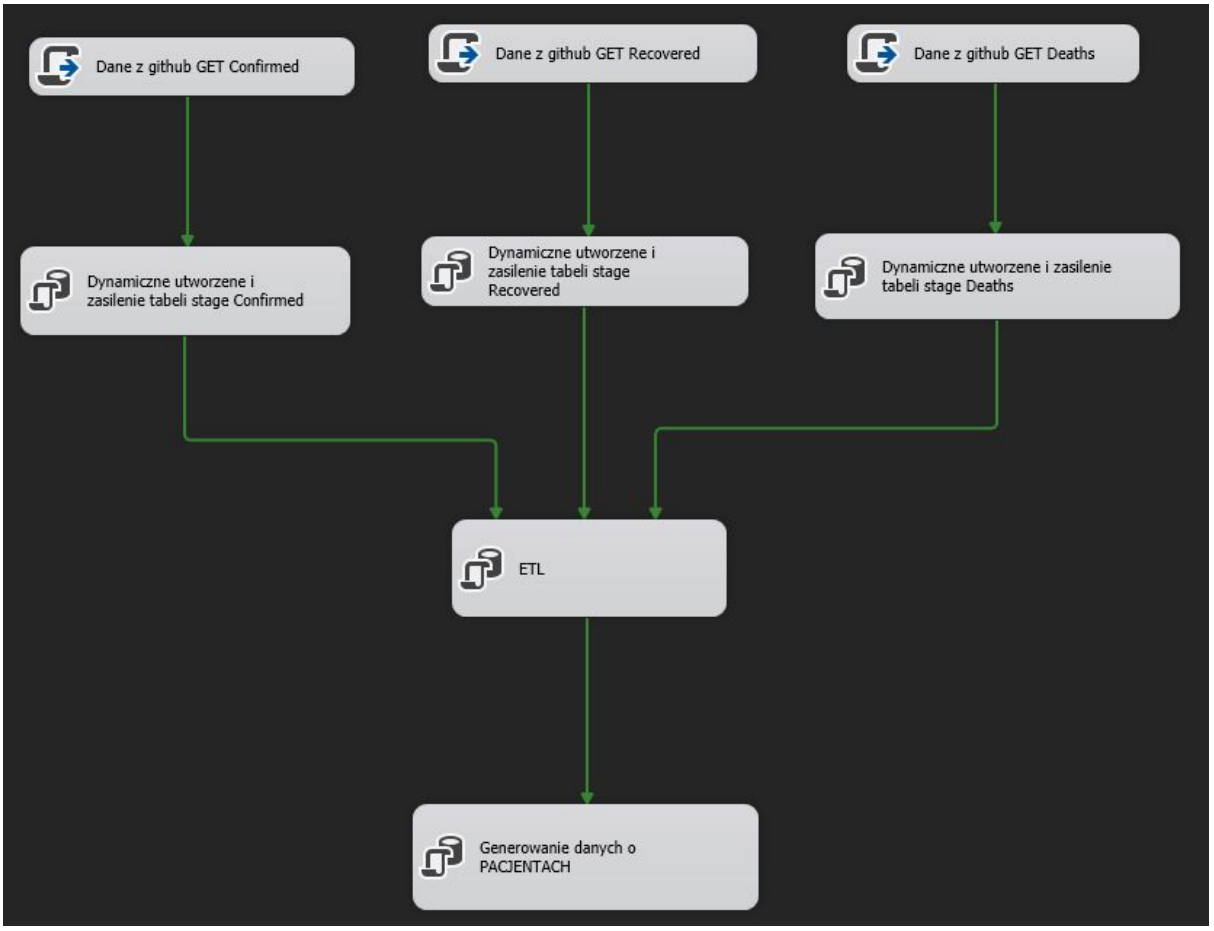


2.3. Proces ETL

Proces inicjacyjny hurtownie danych (*SSIS Package: Init_ETL.dtsx*)



Proces zapewniający inkrementacyjne ładowanie danych (*SSIS Package: Increment_ETL.dtsx*)



Opis procesu ETL:

1. Po pobraniu najnowszych danych na dysk w postaci plików *.csv, dane ładowane są do dynamicznie utworzonych tabel w przestrzeni stage. Za dynamiczne tworzenie tabel i ładowanie do nich danych odpowiada procedura składowana *dbo.utworz_tabele_stage* przyjmująca parametry *nazwa_tabeli, nazwa_pliku_csv, liczba_kolumn_daty.* Liczba kolumn do utworzenia w tabeli obliczana jest za pomocą skryptu (*Dane z github GET (…)*) odpowiadającego za pobranie plików z sieci. Procedura wywoływana jest dla każdego pliku w blokach diagramu procesu z SSIS o nazwie *Dynamiczne utworzenie i zasilenie tabeli stage (...)*

2. Dane które nie wymagają transformacji są bezpośrednio ładowane do przestrzeni stage tabeli faktów dbo.TEMPO_WIRUSA_SUM - dbo.stage_tempo_fact

stage_tempo_fact

Column Name	Data Type	Allow Nulls
FAKT_ID	int	☐
CZAS	nvarchar(50)	☑
GEOGRAFIA	nvarchar(50)	☑
LICZBA_ZAKAZENI_OGO...	int	☑
LICZBA_ZGONOW_OGO...	int	☑
LICZBA_WYLECZONYCH...	int	☑
LICZBA_NOWYCH_ZAKA...	int	☑
LICZBA_NOWYCH_ZGO...	int	☑
LICZBA_NOWYCH_WYLE...	int	☑
LICZBA_ZAKAZONYCH_...	int	☑
DYNAMIKA_ZAKAZEN	decimal(12, 4)	☑
NUMER_KOLEJNY_DNIA	int	☑
		☐

Za podstawowe załadowanie danych z tabel przechowujących surowe dane z plików csv, a taże ich transpozycję odpowiada procedura składowana: *dbo.zasil_tempo_stage_confirmed_recovered_deaths_onthatday*

3. Po załadowaniu podstawowych danych kolejnym etapem jest obliczenie miar.
    a. NUMER_KOLEJNY_DNIA,
    b. LICZBA_NOWYCH_ZAKAZEN_DZIS,
    c. LICZBA_NOWYCH_ZGONOW_DZIS,
    d. LICZBA_NOWYCH_WYLECZONYCH_DZIS,
    e. DYNAMIKA_ZAKAZEN

4. Przyrost (dane dla dni których nie ma w docelowej tabeli faktów TEMPO_WIRUSA_SUM) jest ładowany do tabeli faktów.

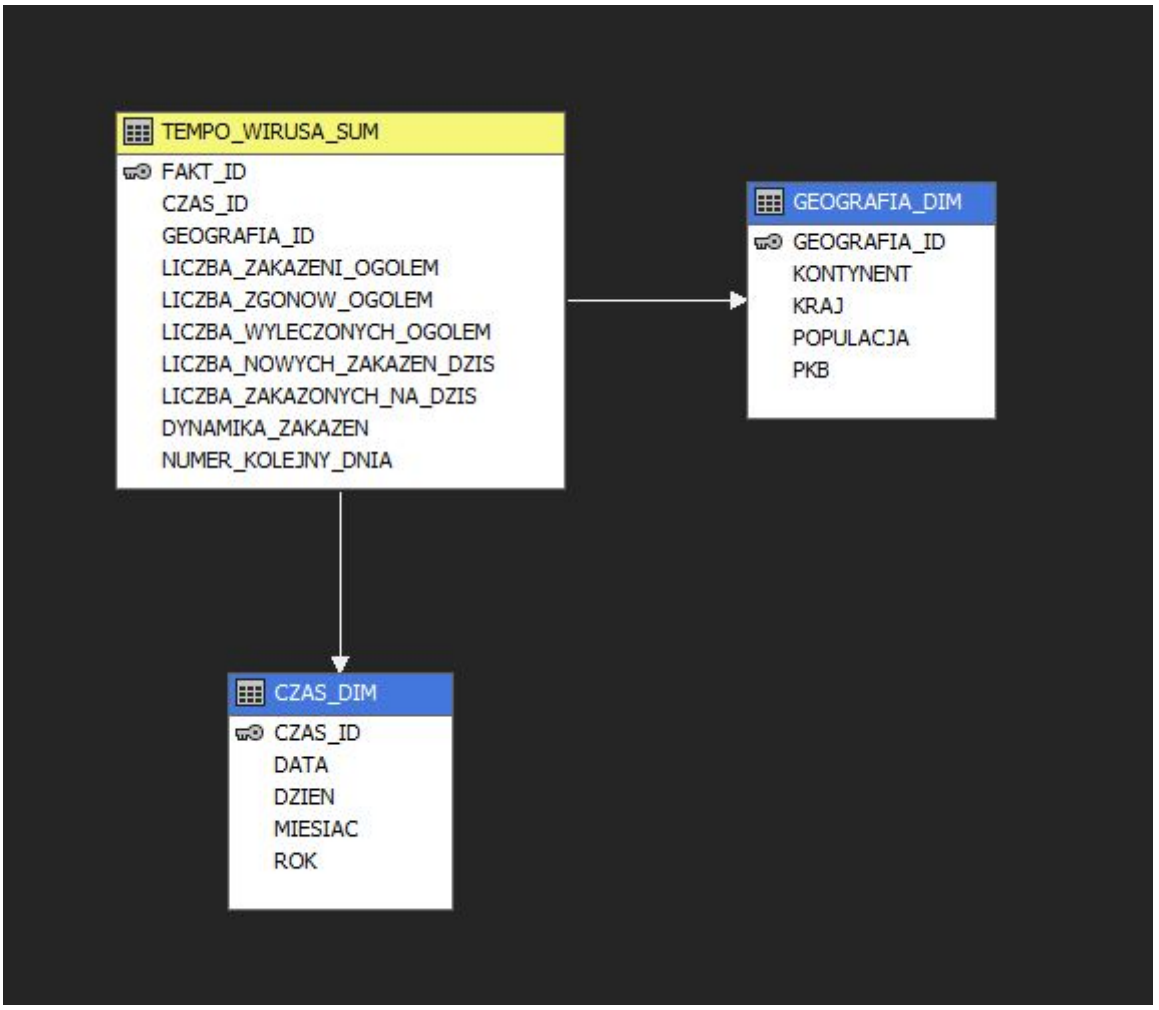
5. Ładowanie danych dotyczących pacjentów odbywają się na podstawie obliczonych wcześniej miar. Za generowanie danych odpowiadają trzy procedury składowane:

- dbo.UtworzNowoZakarzonegoPacjenta - tworzy dane losowe pacjenta oraz wpis w tabeli faktów dbo.ZDARZENIE_PACJENT_DET dotyczące tego zakażenia dla zadanej geografii i czasu.
- dbo.NoweWyzdrowienia - tworzy grupę zdarzeń - wyzdrowień, dla zadanej geografii i czasu,
- dbo.NoweSmierci - tworzy grupę zdarzeń - śmierci, dla zadanej geografii i czasu.

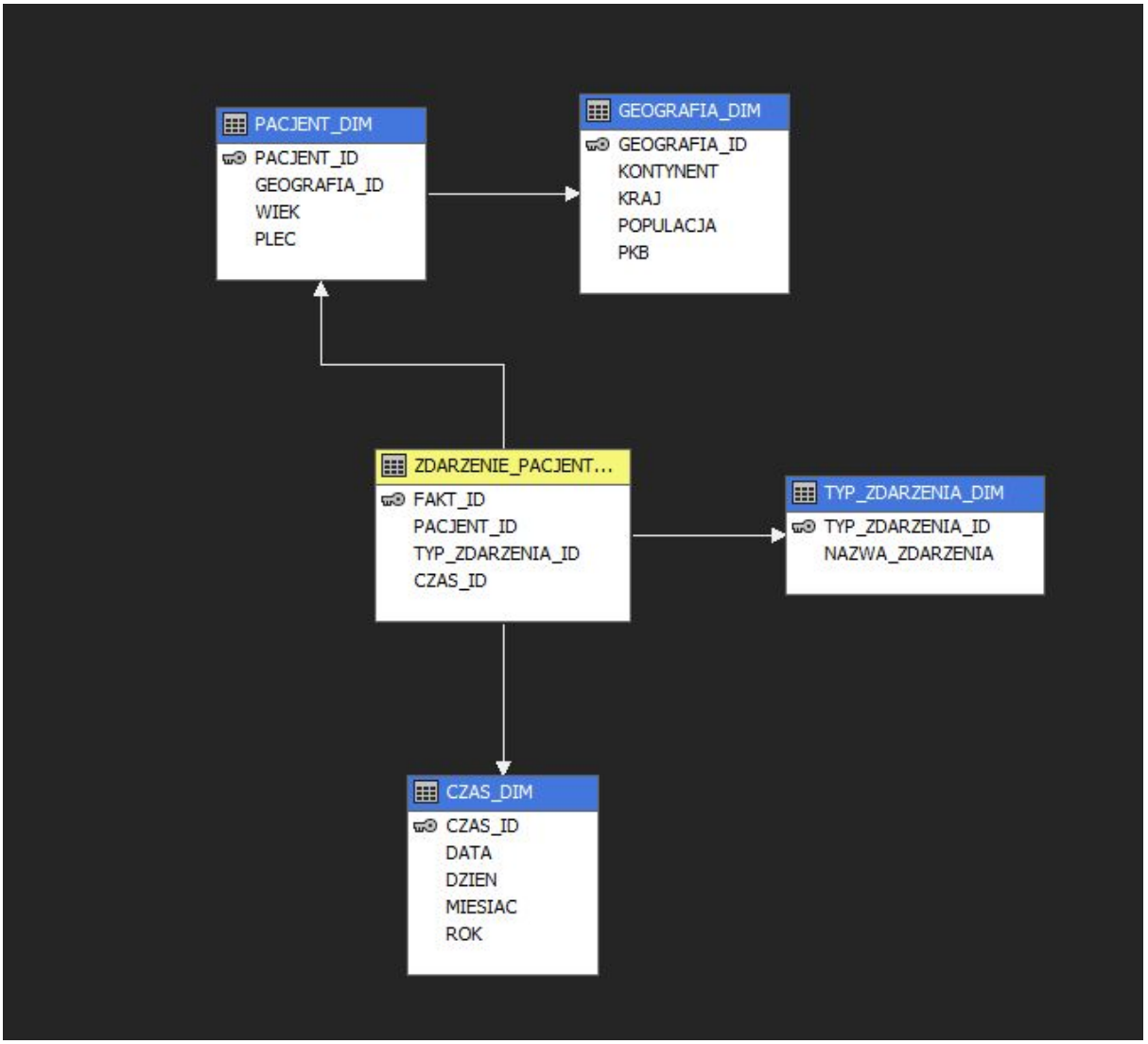
Ładowanie odbywa się w pętlach przechodzących po każdym wierszu tabeli *stage_tempo_fact* i korzystających z liczby nowych zakażeń,wyleczonych, zgonów d danym dniu.

2.4. Model wymiarowy - Analysis Services (model kostki)

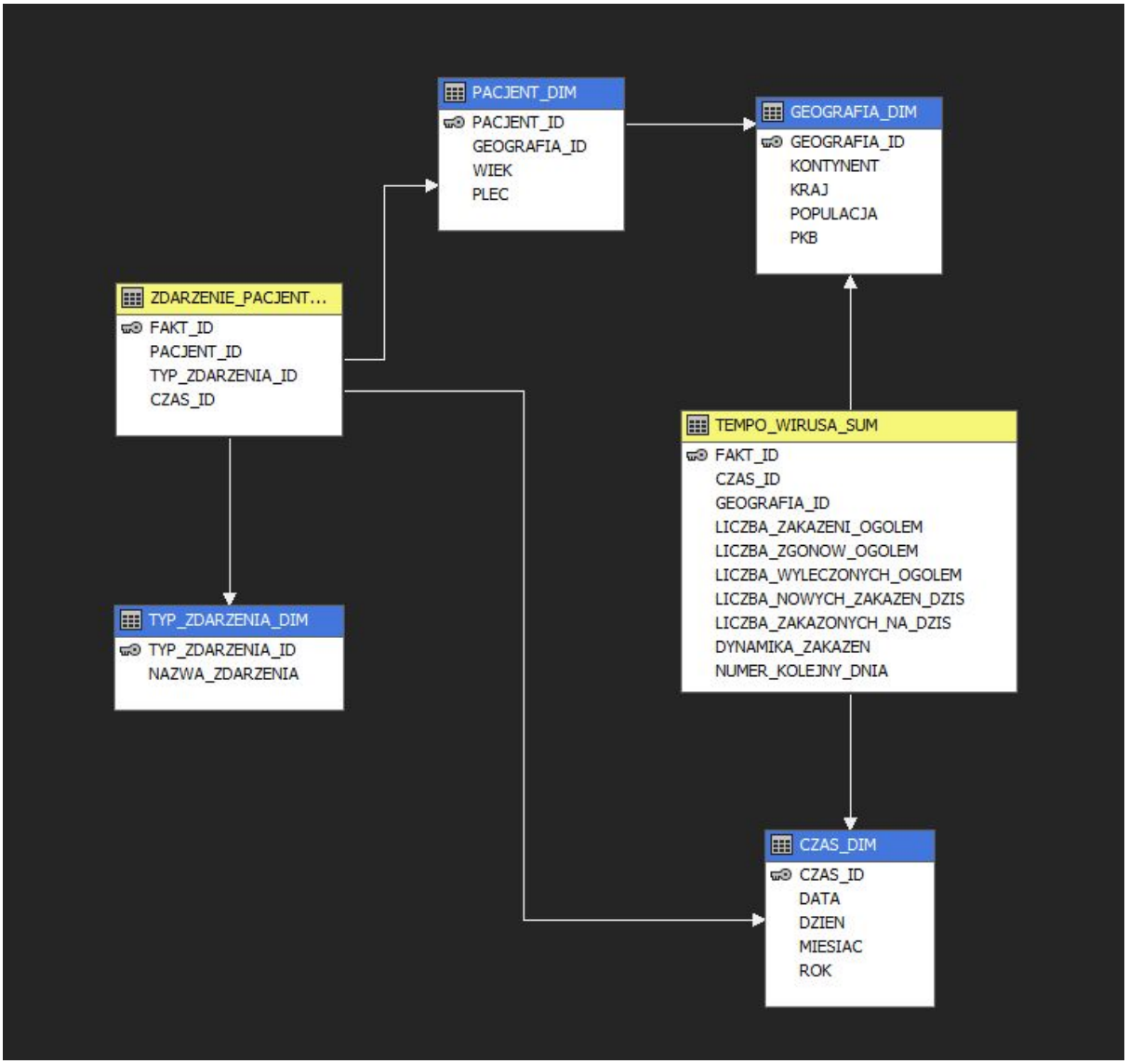
2.5. Kostka TEMPO WIRUSA



2.6. Kostka PACJENT

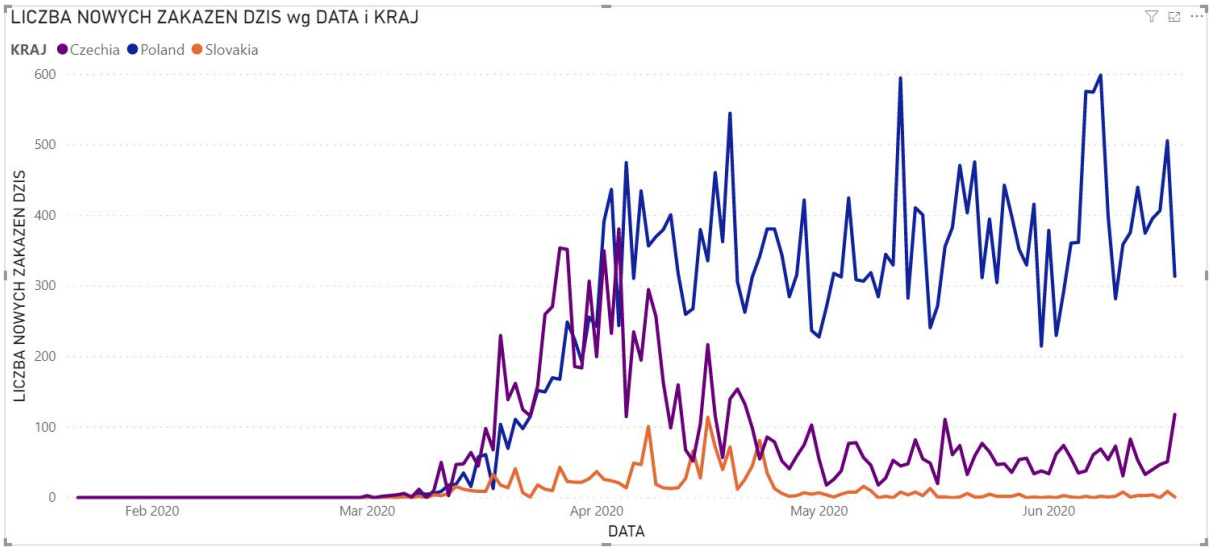


2.7. Koska TEMPO WIRUSA + PACJENT

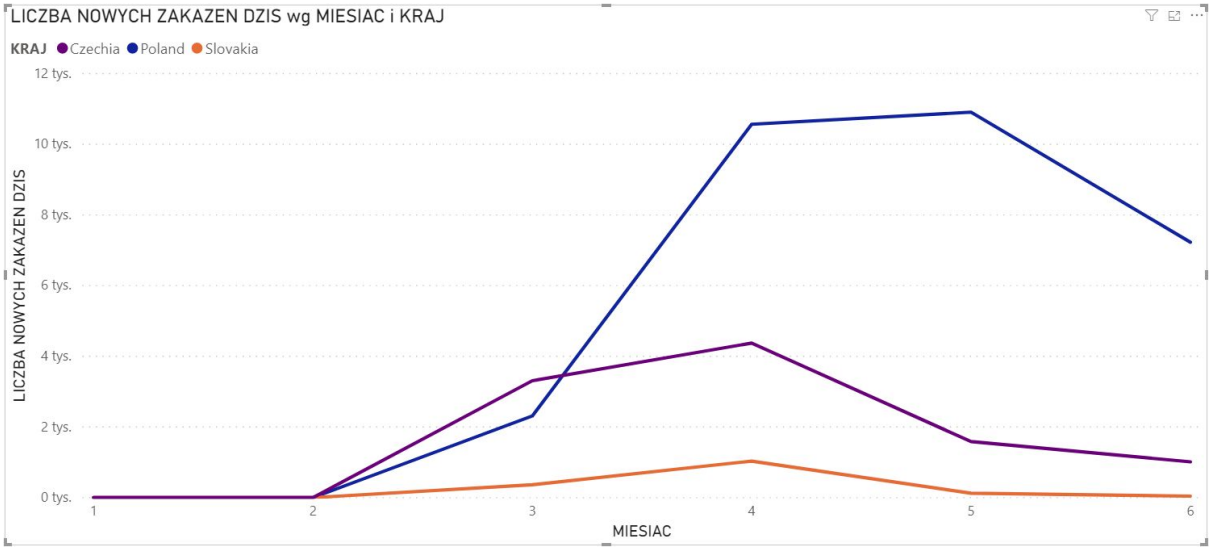


3. Przykładowe raporty

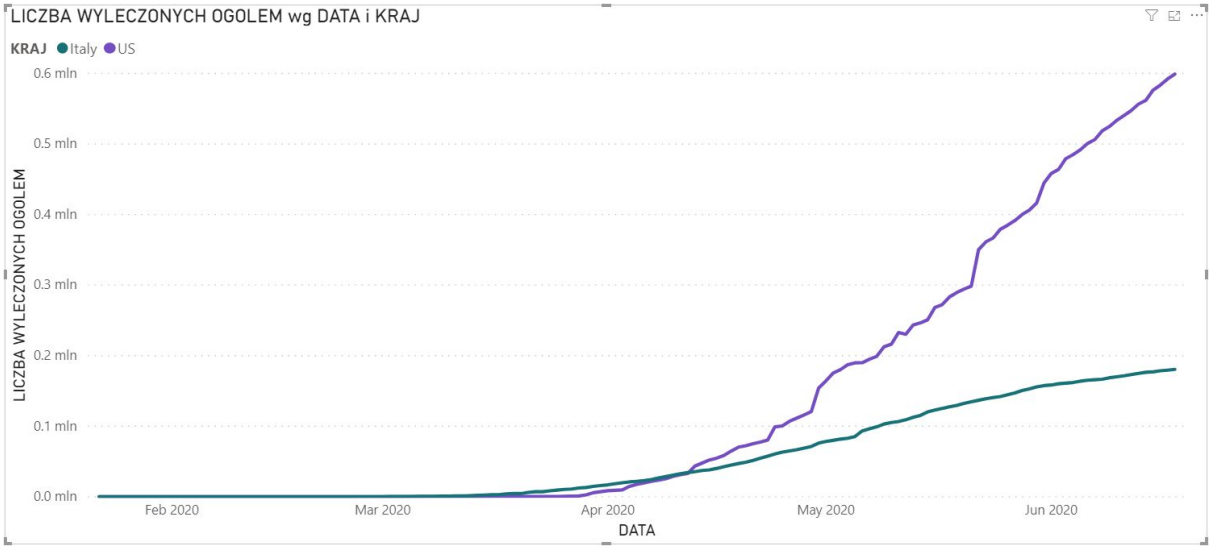
Nowe przypadki zachorowań w Polsce, Słowacji i Czechach (granulacja dzienna)



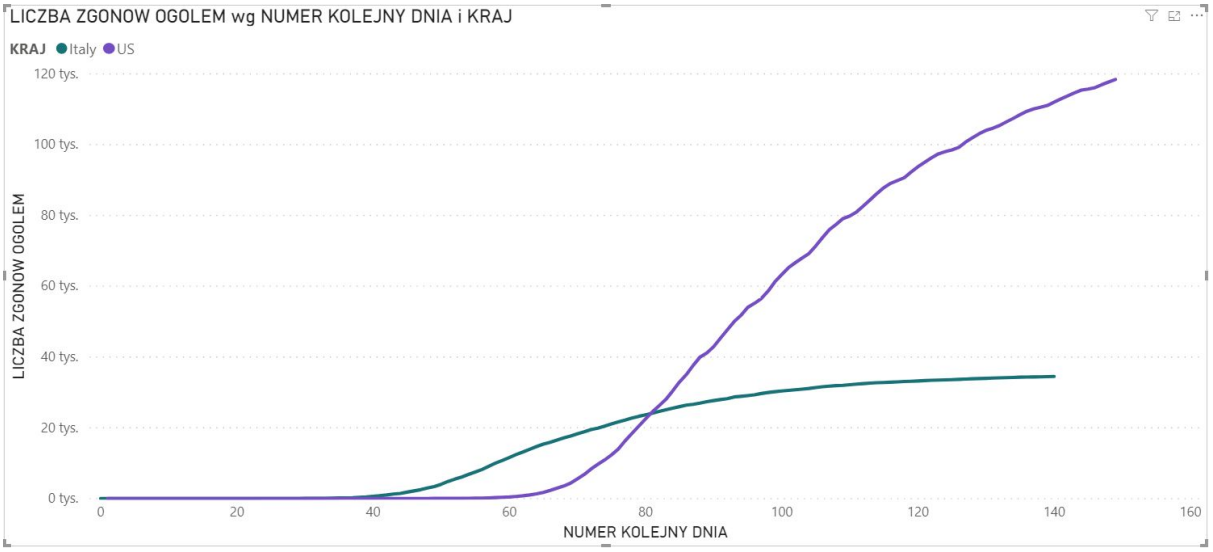
Nowe przypadki zachorowań w Polsce, Słowacji i Czechach (granulacja miesięczna)



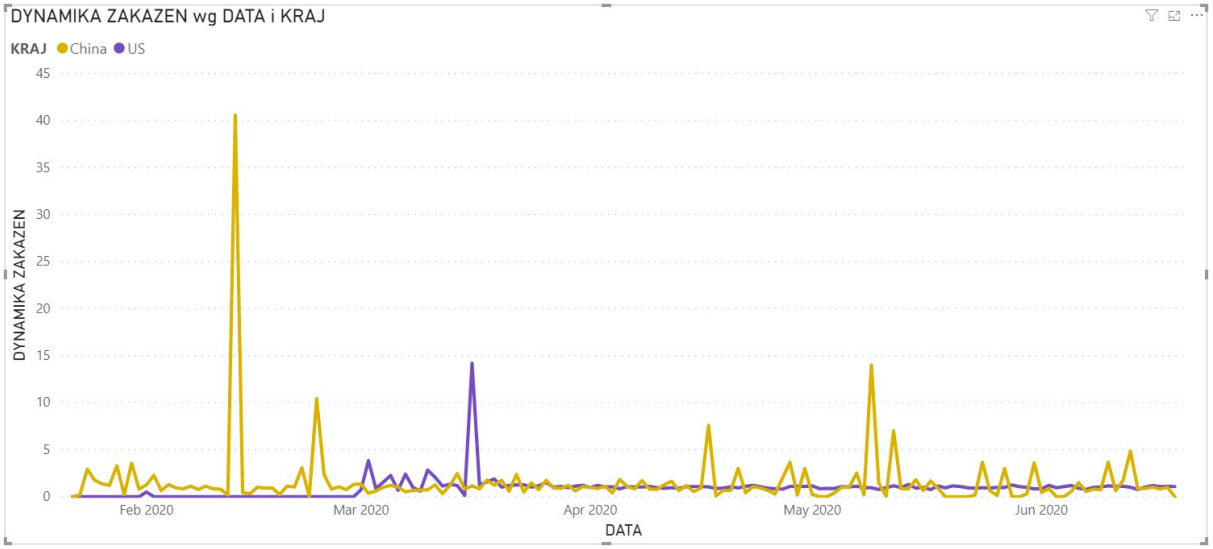
Liczba wyleczonych przypadków ogółem w USA i we Włoszech (granulacja dzienna)



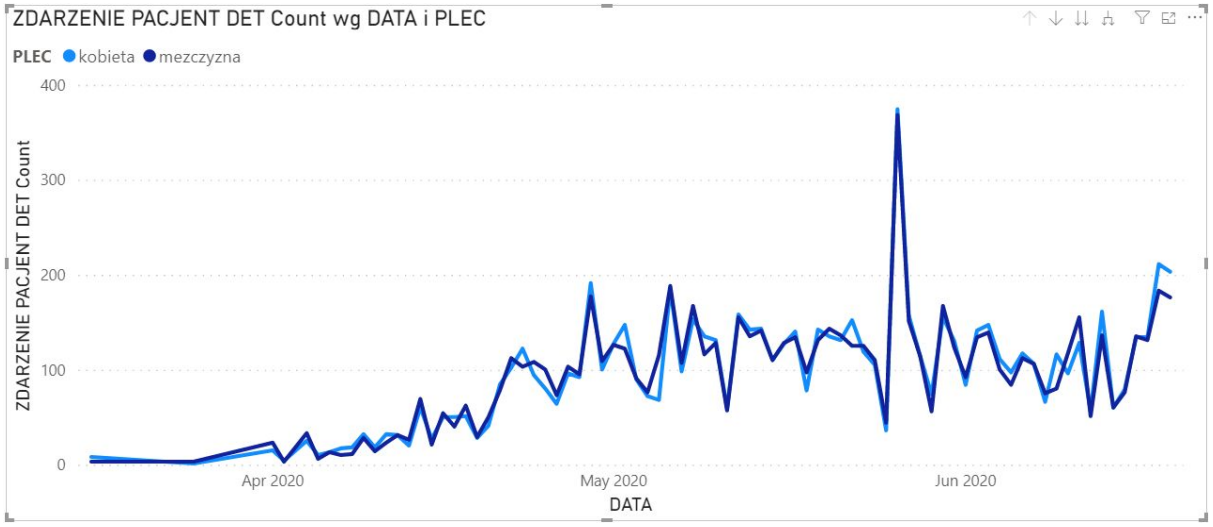
Liczba zgonów ogółem w USA i we Włoszech (w zależności od numeru kolejnego dnia od pierwszego zakażenia w danym kraju)



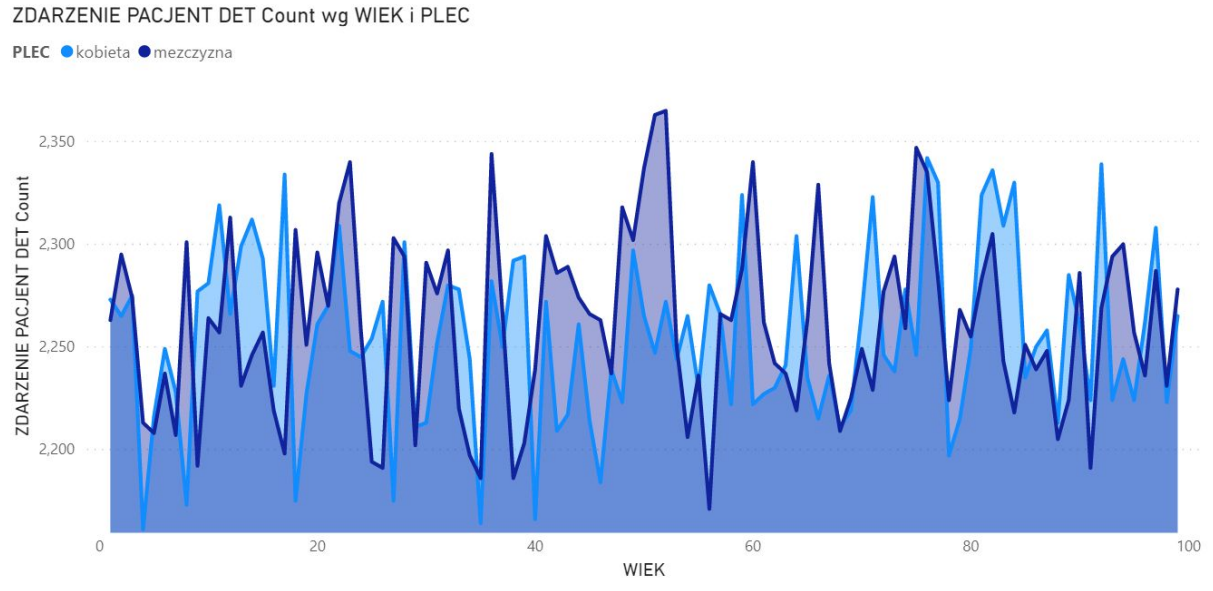
Dynamika zakażeń w Chinach i USA



Liczba wyzdrowień w Polsce w granulacji dziennej w zależności od płci



Liczba zgonów ogółem w zależności od wieku i płci



4. Repozytorium projektu

Kody źródłowe oraz projekty SSIS i SSMS, przykładowy plik PowerBI zostały umieszczone w repozytorium kodów pod adresem:
https://github.com/ddavid09/HurtowniaDanychCovid