**Wojskowa Akademia Techniczna**

**Hurtownie Danych**

**Sprawozdanie z projektu**

Prowadzący: dr inż. Marcin Mazurek

Wykonawcy: Dominik Marchewka

Piotr Skomorowski

Grupa: I7B1S1

1. Zadanie projektowe

Wykorzystując dane o rozprzestrzenianiu się w koronowirusa zbudować hurtownię danych umożliwiającą analizę tempa rozprzestrzeniania się wirusa:

**Wymiary analizy:**

* Geografia – kraj (atrybuty: populacja, GDP) – kontynent,
* Czas – dzień – miesiąc – rok,
* Czas od pierwszej detekcji (numer kolejny dnia),
* Pacjent (wobec braku danych należy je wygenerować losowo lub próbki danych z Kaggle) (wiek, płeć).

**Miary:**

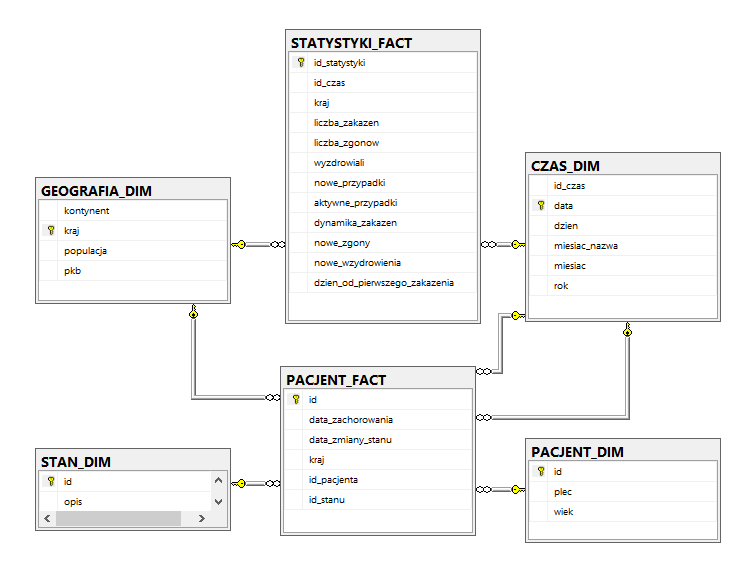
* liczba zakażeń (w okresie),
* liczba zgonów (w okresie),
* liczba pacjentów wyleczonych (w okresie),
* liczba nowych przypadków zakażeń (granulacja dzienna),
* liczba pacjentów zakażonych (stan na dzień),
* dynamika zakażeń – liczba nowych przypadków / liczba pacjentów zakażonych w dniu poprzedzającym.

1. Opis założeń biznesowych

Hurtownia danych ma umożliwiać przyrostowe dodawanie danych na temat rozprzestrzeniania się choroby COVID – 19. Zostanie ona utworzona za pomocą SQL Server 2019. Hurtownia powinna przechowywać informację na temat liczby zakażonych, zgonów, wyzdrowiałych, dynamiki zakażeń, liczny dni od zakażenia oraz informacje na temat pojedynczego pacjenta, w zależności od geografii (kraj, kontynent) oraz daty. Proces ETL powinien być uruchamiany przez administratora. Hurtownia ma umożliwiać tworzenie raportów na temat pandemii koronawirusa w programie PoewrBI.

1. Model bazy danych

Do realizacji zadani utworzono poniższą bazę danych:

Zawiera ona następujące tabele wymiarów:

* CZAS\_DIM – wymiar czasu,
* GEOGRAFIA\_DIM – wymiar obszarów geograficznych,
* STAN\_DIM – wymiar stanów pacjenta,
* PACJENT\_DIM – wymiar pacjenta,

oraz tabele faktów:

* STATYSTYKI\_FACT – tabela statystyk rozprzestrzeniania się wirusa,
* PACJENT\_FACT – tabela faktów pojedynczego pacjent.

1. Proces ETL

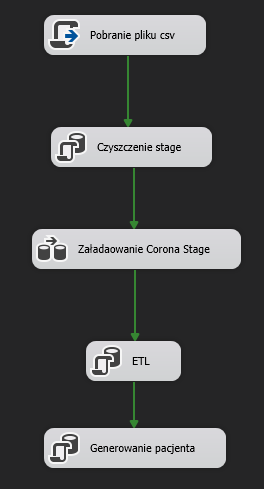
Dane w postaci pliku csv, na temat pandemii koronawirusa pochodzą ze strony:

<https://raw.githubusercontent.com/datasets/covid-19/master/data/countries-aggregated.csv>

Zawierają one następujące kolumny danych:

* Date - data
* Country - kraj
* Confirmed – ilość potwierdzonych przypadków danego dnia
* Recovered – ilość wyzdrowiałych osób danego dnia
* Deaths – ilość zgonów danego dnia

Implementacja procesu ETL:



Obraz zawierający zrzut ekranu

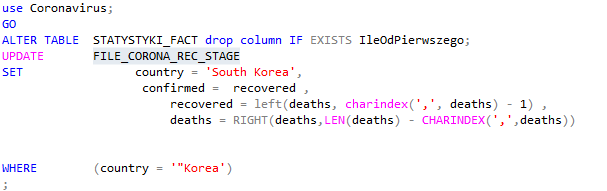
Opis wygenerowany automatycznieDane są automatycznie pobierane za pomocą skryptu, napisanego w języku C#. Poniżej zamieszczono jego kod:

Następnie czyszczona jest tablica FILE\_CORONA\_REC\_STAGE, za pomocą skryptu w SQL:

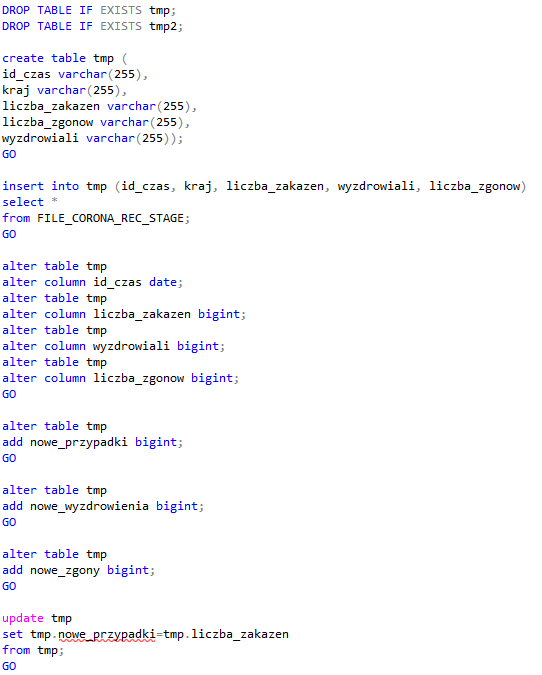


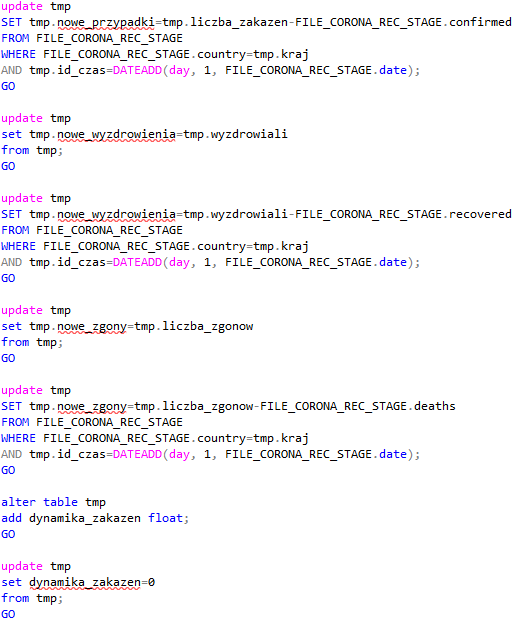
Następnie plik csv jest importowany do pustej tabeli STAGE, przy pomocy Data Flow Task.

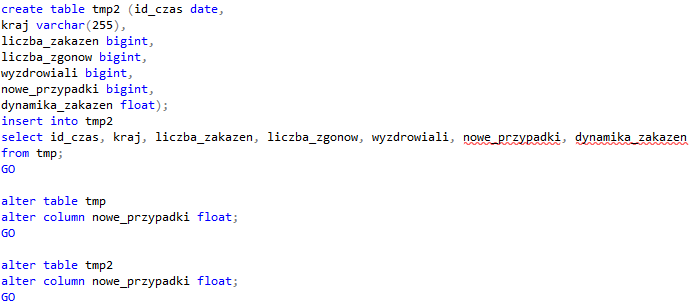
W następnej kolejności wykonywany jest skrypt SQL pod nazwą ETL. W pierwszej kolejności usuwana jest tabela IleOdPierwszego i w tabeli FILE\_CORONA\_REC\_STAGE, zamieniana jest „Korea, South” na ‘South Korea’.

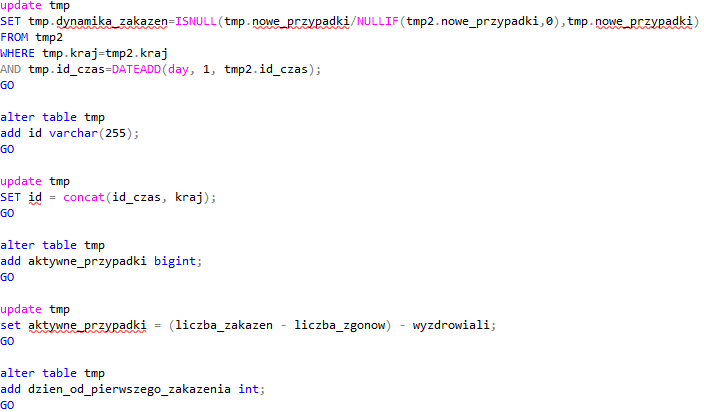


W następnej kolejności tworzone są tabele pomocnicze, gdzie konwertowane są dane oraz wyliczane potrzebne statystyki.

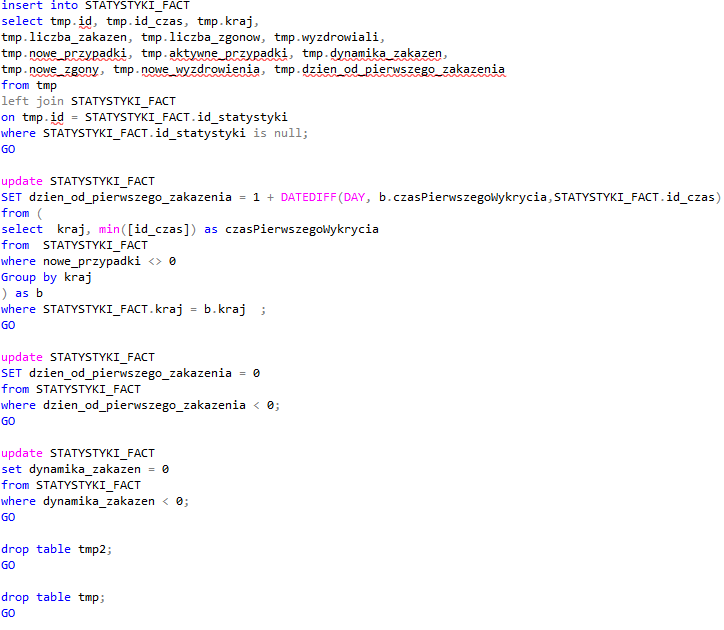




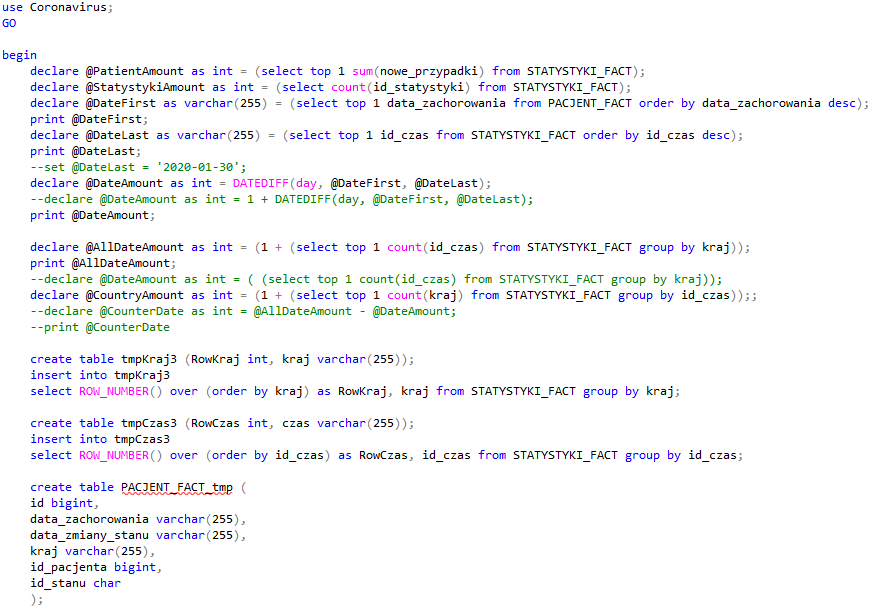


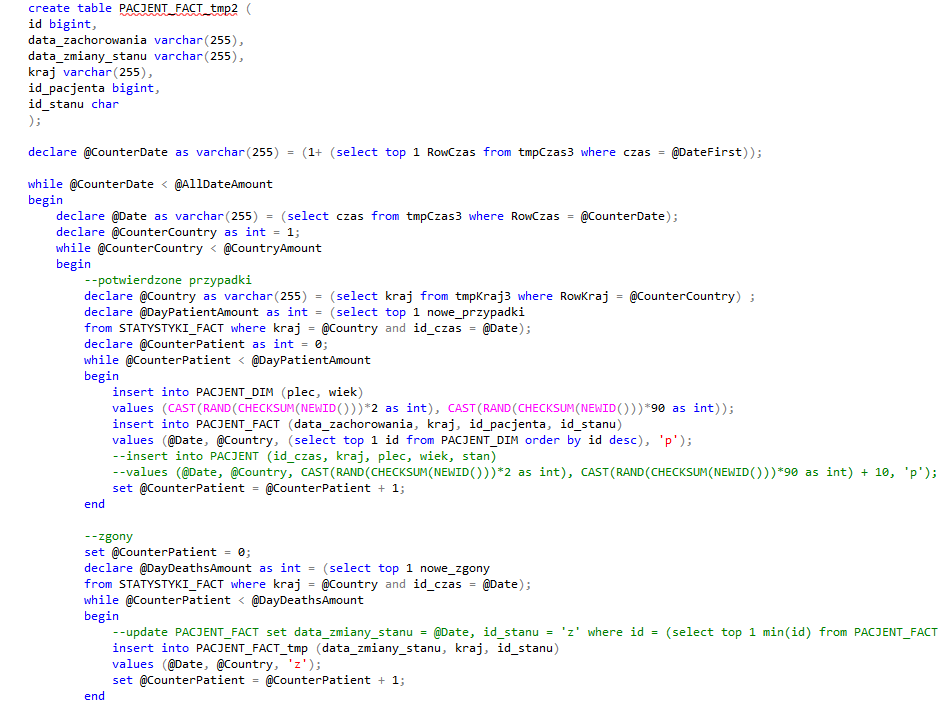


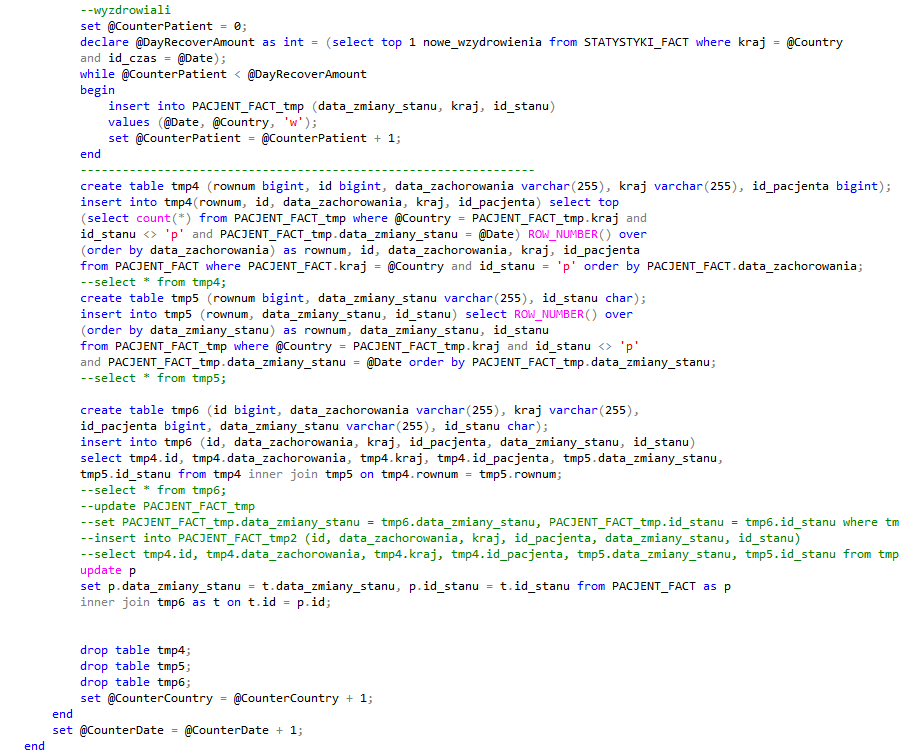
Ładownie danych do tabeli STATYSTKI\_FACT i korygowanie ujemnych danych.

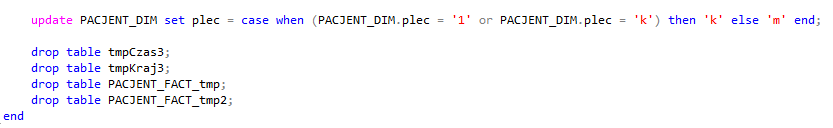


Po wykonaniu skryptu ETL, uruchamiany jest skrypt generowania pacjenta. Pacjent jest generowany w trzech pętlach (daty, kraju, ilości pacjentów). Początkowo do kolumny PACJENT\_FACT generowani są nowi pacjenci, wszyscy z id\_stanu = p. Następnie przy pomocy tabel pomocniczych edytowana jest tabela PACJENT\_FACT poprzez zmienienie id\_stanu i data\_zmiany\_stanu na podstawie nowych wyzdrowiałych oraz nowych zgonów.

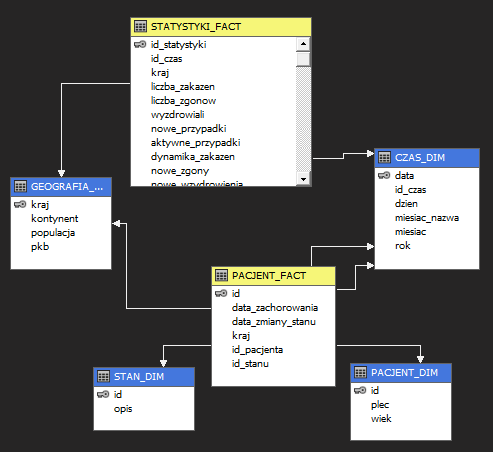








1. Model kostki

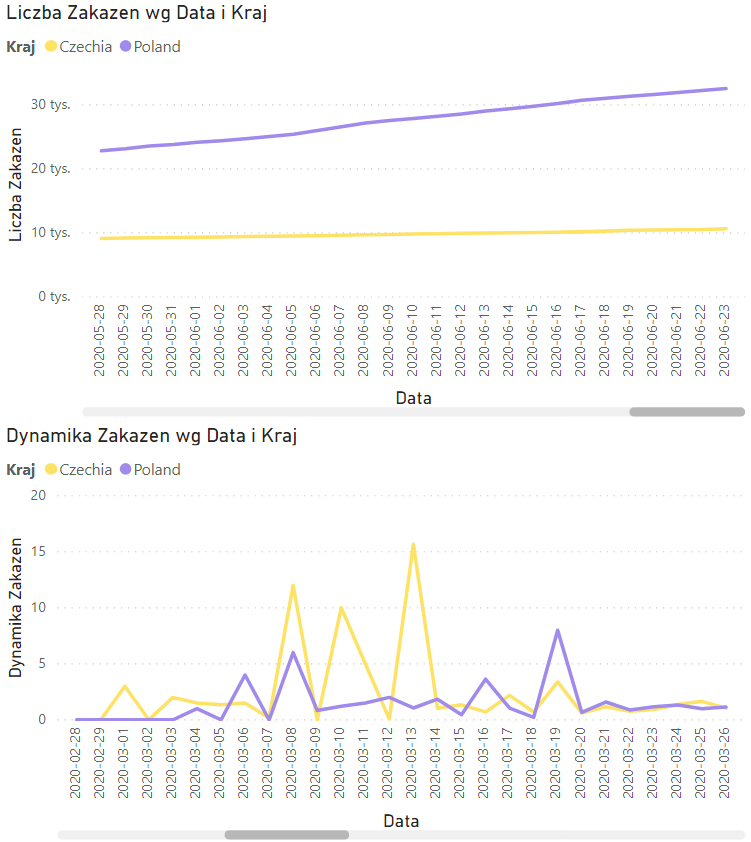


Miary:

* STATYSTKI\_FACT
* PACJENT\_FACT

Wymiary:

* GEOGRAFIA\_DIM
* STAN\_DIM
* PACJENT\_DIM
* CZAS\_DIM

1. Przykładowe raporty w PowerBI

