# Opracowanie modelu pojęciowego bazy danych przestrzennych MSP w województwie kujawsko-pomorskim

Projekt „Usytuowanie na poziomie samorządów lokalnych instrumentów wsparcia dla MŚP, działających w oparciu o model wielopoziomowego zarządzania regionem”
realizowany przez konsorcjum w składzie: Województwo Kujawsko-Pomorskie, Szkoła Główna Handlowa w Warszawie i Uniwersytet Mikołaja Kopernika w Toruniu";
finansowany z Narodowego Centrum Badań i Rozwoju
w ramach Programu Strategicznego Gospostrateg –
„Społeczny i Gospodarczy Rozwój Polski
w warunkach globalizujących się rynków”
Nr umowy GOSPOSTRATEG1/385453/3/NCBR/2018

# Rozdział 1

Struktura bazy danych: lista tabel i widoków w podziale na schematy, z adnotacją o sposobie pozyskania danych agregatach i aktualności. Uwagi wyróżnione kolorem

## 1.1 Schemat BDL

Zawiera wybrane zakresy danych udostępnianych przez GUS w Banku Danych Lokalnych

**Tabela BEZROBOCIE**

Dane o liczbie bezrobotnych zarejestrowanych, z BDL, dla lat 2017, 2018, 2019.

**Tabela DOCHODYJA**

Dochody jednostek administracyjnych, w tym dochody własne, w tym z podatku dochodowego od osób fizycznych i osób prawnych i z podatków ustalanych i pobieranych na podstawie odrębnych ustaw. Dane dla lat 2017 i 2018, dla gmin, powiatów oraz dla województwa.

**Tabela LLUDNOSCI**

Liczba ludności ogółem, w wieku przedprodukcyjnym, produkcyjnym i poprodukcyjnym. Dane dla lat 2017 i 2018, dla gmin, powiatów oraz dla województwa.

**Tabela LLUDNOSCI_CZESCIGMIN**

Dane dotyczące liczby ludności wg charakterystyki tabeli LLUDNOSCI, dotyczące gmin miejsko-wiejskich w podziale na część miejską i część wiejską, uzupełnione o TERYT gminy jako całości.

**Tabela REGON**

Dane o liczbie podmiotów w rejestrze REGON. Bez osób prowadzących gospodarstwa indywidualne w rolnictwie. Podmioty klasyfikowane według kryterium liczby pracujących. Dane dla miejscowości statystycznych z rejestru Regon podawane są wg: - adresu zamieszkania dla osób fizycznych z krajowym adresem zamieszkania, - adresu siedziby dla pozostałych jednostek tj. osób fizycznych z zagranicznym adresem zamieszkania, osób prawnych i jednostek organizacyjnych niemających osobowości prawnej oraz jednostek lokalnych. Dane dla roku 2017 i 2018 dla gmin.

**Tabela REGON_CZESCIGMIN**

Dane o liczbie podmiotów w rejestrze REGON w podziale na liczbę pracujących dla części miejskiej i wiejskiej gmin miejsko-wiejskich, uzupełnione o TERYT części gminy i gminy jako całości.

**Tabela SZKOLYZAW_CZESCIGMIN**

Dane dotyczące szkół zawodowych - liczba szkół i absolwentów w częściach wiejskich i miejskich gmin miejsko-wiejskich. Dane dla 2017 i 2018.

**Tabela SZKOLYZAWODOWE**

Dane dotyczące szkół zawodowych - liczba szkół i absolwentów w poszczególnych gminach w latach 2017 i 2018.

**Tabela WYDATKIJA**

Wydatki ogółem i wydatki majątkowe gmin w latach 2017 i 2018.

**Tabela ZATRUDNIENIE**

Struktura zatrudnienia: liczba osób zatrudnionych w danym dziale gospodarki i procent całości. W podziale na działy: rolnictwo, handel, przemysł, finanse i pozostałe. Dane dla powiatów, dla roku 2018

W miarę możliwości uzupełnić o rok 2017.

**Tabela ZOBOWIAZANIAJA**

Zobowiązania jednostki administracyjnej ogółem [PLN] i w przeliczeniu na mieszkańca [PLN/osobę], dane dla powiatów i gmin dla roku 2017, 2018, 2019.

Uwaga: zobowiązania na mieszkańca dla roku 2019 wyliczone: ogółem w 2019/liczba ludności 2018. W BDL nie ma jeszcze liczby ludności dla 2019

**Widok BEZROBOCIE_GMINY2017**

Widok na tabeli BEZROBOCIE, bezrobocie rejestrowane w gminach w roku 2017.

**Widok BEZROBOCIE_GMINY2018**

Widok na tabeli BEZROBOCIE, bezrobocie rejestrowane w gminach w roku 2018.

**Widok BEZROBOCIE_GMINY2019**

Widok na tabeli BEZROBOCIE, bezrobocie rejestrowane w gminach w roku 2019.

Gdy uzyskamy dane o liczbie ludności w roku 2019 warto uzupełnić o stopę bezrobocia

**Widok BEZROBOCIE_POWIATY2017**

Widok na tabeli BEZROBOCIE, bezrobocie rejestrowane w powiatach w roku 2017.

**Widok BEZROBOCIE_POWIATY2018**

Widok na tabeli BEZROBOCIE, bezrobocie rejestrowane w powiatach w roku 2018.

**Widok BEZROBOCIE_POWIATY2019**

Widok na tabeli BEZROBOCIE, bezrobocie rejestrowane w powiatach w roku 2019.

Gdy uzyskamy dane o liczbie ludności w roku 2019 warto uzupełnić o stopę bezrobocia

**Widok REGON_2017**

Widok na tabeli REGON, dane o podmiotach zarejestrowanych w REGON w roku 2017.

**Widok REGON_2018**

Widok na tabeli REGON, dane o podmiotach zarejestrowanych w REGON w roku 2018.

**Widok REGON_POW2017**

Widok na tabeli REGON, dane o podmiotach zarejestrowanych w REGON w roku 2017 zagregowane do powiatów.

**Widok REGON_POW2018**

Widok na tabeli REGON, dane o podmiotach zarejestrowanych w REGON w roku 2018 zagregowane do powiatów.

**Widok ZATRUDNIENIE_OSOBY2018**

Widok na tabeli ZATRUDNIENIE, agreguje dane o liczbie osób zatrudnionych w danym dziale gospodarki w roku 2018 do jednego rekordu per powiat.

**Widok ZATRUDNIENIE_PROCENT2018**

Widok na tabeli ZATRUDNIENIE, agreguje procentowa strukturę zatrudnienia w działach gospodarki do jednego rekordu per powiat. Dane dla 2018r.

## 1.2 Schemat BDOT10k

Baza Danych Obiektów Topograficznych o rozdzielczości odpowiadającej mapie topograficznej w skali 1:10 000 przekształcona do postaci ciągłej dla województwa.

Tabele:

OT_ADJA_A,
OT_ADMS_A,
OT_ADMS_P,
OT_BUBD_A,
OT_BUCM_A,
OT_BUHD_A,
OT_BUHD_L,
OT_BUHD_P,
OT_BUIB_A,
OT_BUIB_L,
OT_BUIN_L,
OT_BUIT_A,
OT_BUIT_P,
OT_BUSP_A,
OT_BUSP_L,
OT_BUTR_L,
OT_BUTR_P,
OT_BUUO_L,
OT_BUWT_A,
OT_BUWT_P,
OT_BUZM_L,
OT_BUZT_A,
OT_BUZT_P,
OT_Ciek,
OT_Elektrownia,
OT_Kopalnia,
OT_KUHO_A,
OT_KUHU_A,
OT_KUHU_P,
OT_KUIK_A,
OT_KUKO_A,
OT_KUKO_P,
OT_KUMN_A,
OT_KUOS_A,
OT_KUOZ_A,
OT_KUPG_A,
OT_KUPG_P,
OT_KUSC_A,
OT_KUSK_A,
OT_KUZA_A,
OT_LiniaKolejowa,
OT_Lotnisko,
OT_OIKM_A,
OT_OIKM_L,
OT_OIKM_P,
OT_OIMK_A,
OT_OIOR_A,
OT_OIOR_L,
OT_OIOR_P,
OT_OIPR_L,
OT_OIPR_P,
OT_OISZ_A,
OT_Port,
OT_PTGN_A,
OT_PTKM_A,
OT_PTLZ_A,
OT_PTNZ_A,
OT_PTPL_A,
OT_PTRK_A,
OT_PTSO_A,
OT_PTTR_A,
OT_PTUT_A,
OT_PTWP_A,
OT_PTWZ_A,
OT_PTZB_A,
OT_SKDR_L,
OT_SKJZ_L,
OT_SKPP_L,
OT_SKRP_L,
OT_SKRW_P,
OT_SKTR_L,
OT_SULN_L,
OT_SUPR_L,
OT_SWKN_L,
OT_SWRM_L,
OT_SWRS_L,
OT_SzlakDrogowy,
OT_TCON_A,
OT_TCPK_A,
OT_TCPN_A,
OT_TCRZ_A,
OT_Uica,
OT_WezelKolejowy,
OT_ZbiornikWodny,

oraz słowniki i tabele intersekcji.

## 1.3 Schemat PODATKI

Dane o podatnikach zagregowane do gmin uzyskane z Krajowej Administracji Skarbowej

**Tabela CIT8**

Dane pozyskane z deklaracji podatkowych CIT8 zagregowane do gmin. Dane dla roku 2016, 2017 i 2018.

**Tabela PIT28A**

Dane pozyskane z deklaracji podatkowych PIT28A zagregowane do gmin. Dane dla roku 2016, 2017 i 2018.

**Tabela PIT28B**

Dane pozyskane z deklaracji podatkowych PIT28B zagregowane do gmin. Dane dla roku 2016, 2017 i 2018.

**Tabela PIT36**

Dane pozyskane z deklaracji podatkowych PIT36 zagregowane do gmin. Dane dla roku 2016, 2017 i 2018..

**Tabela PIT36L**

Dane pozyskane z deklaracji podatkowych PIT36L zagregowane do gmin. Dane dla roku 2016, 2017 i 2018

**Tabela VAT**

Dane pozyskane z deklaracji podatkowych VAT zagregowane do gmin. Dane dla roku 2016, 2017, 2018 i pierwszych 4 miesięcy roku 2019.

## 1.4 Schemat PUBLIC

**Tabela SEKCJA_KATEGORIA**

Mapowanie sekcji Polskiej Klasyfikacji Działalności na kategorie działalności z grubsza i bardziej szczegółowo. Tabela pomocnicza

**Tabela TERYT_KOD**

Tabela pomocnicza: Jednostki administracyjne wg kodu TERYT, kod i nazwa w danych US, kod i nazwa w danych BDL, nazwa w danych ZUS, oznaczenie typu gminy.

## 1.5 Schemat PWC

Dane dotyczące projektów z analizy społeczno-gospodarczej, którą przygotowywało PwC i TARR. Excele zaimportowane bez ingerencji w strukturę, dodany TERYT jednostki administracyjnej, tam gdzie było to możliwe.

## 1.6 Schemat ZUS

Dane uzyskane z Zakładu Ubezpieczeń Społecznych zagregowane do gmin z zachowaniem zasad anonimizacji.
Wielkość przedsiębiorstwa w przedziałach: samozatrudnieni, od 1 do 9 osób, od 10 do 49 osób, od 50 do 249 osób, ponad 250 osób.
Wiek w przedziałach: poniżej 24 lat, od 25 do 34 lat, od 35 do 44 lat, od 45 do 54 lat, od 55 do 64 lat, powyżej 65 lat
Dane wymagają uzupełnienia,są niekompletne lub nadmiernie zagregowane, dla gmin o nieunikalnych nazwach: Aleksandrów Kujawski, Brodnica, Chełmno, Chełmża, Golub-Dobrzyń, Inowrocław, Kowal, Lipno, Radziejów, Rypin, Rogowo, Włocławek przekazano tylko po jednej jednostce.

**Tabela PLATNICY**

Liczba płatników płacących składki emerytalne w grudniu danego roku w gminach województwa kujawsko-pomorskiego według sekcji PKD i wielkości przedsiębiorstwa. Dane dla roku 2017 i 2018.

**Tabela PLATNICY_PODWOJNEGMINY**

Liczba płatników z tabeli PLATNICY w gminach o podwójnych nazwach, oszacowana na podstawie danych podatkowych, dane dla roku 2018.

**Tabela PLATNICYSAM**

Liczba płatników płacących składki emerytalne sam za siebie w grudniu danego roku w gminach województwa kujawsko-pomorskiego według sekcji PKD, płci i wieku. Dane dla lat 2017 i 2018.

**Tabela PODSTAWY**

Podstawa składki emerytalnej w danym roku w podziale na sekcje PKD, wielkość płatnika, wiek i płeć zatrudnionych. Dane dla lat 2017 i 2018.

**Tabela PODSTAWYSAM**

Podstawa składki emerytalnej samozatrudnionych w danym roku w podziale na sekcje PKD, wiek i płeć zatrudnionych. Dane dla lat 2017 i 2018.

**Tabela RAPORT**

Podmioty zarejestrowane lub wyrejestrowane z ZUS w danym roku w podziale na sekcje PKD i wielkość płatnika. Dane dla lat 2017 i 2018.

**Tabela UBEZPIECZENI**

Liczba osób ubezpieczonych w ZUS w danym roku w podziale na sekcje PKD, wielkość podmiotu ubezpieczającego, wiek i płeć zatrudnionych.

**Widok PLATNICY_SEKCJA2017**

Widok na tabeli PLATNICY. Suma liczby płatników w podziale na sekcje PKD dla roku 2017 (agregacja po wielkości płatnika).

**Widok PLATNICY_SEKCJA2018**

Widok na tabeli PLATNICY. Suma liczby płatników w podziale na sekcje PKD dla roku 2018 (agregacja po wielkości płatnika).

**Widok PLATNICY_WIELKOSCPL2017**

Widok na tabeli PLATNICY. Suma liczby płatników w podziale na wielkość płatnika dla roku 2017 (agregacja po sekcji).

**Widok PLATNICY_WIELKOSCPL2018**

Widok na tabeli PLATNICY. Suma liczby płatników w podziale na wielkość płatnika dla roku 2018 (agregacja po sekcji).

**Widok PLATNICYSAM_PLEC2017**

Widok na tabeli PLATNICYSAM. Suma liczby płatników sam za siebie w podziale na płeć dla roku 2017 (agregacja po sekcji PKD i wieku).

**Widok PLATNICYSAM_PLEC2018**

Widok na tabeli PLATNICYSAM. Suma liczby płatników sam za siebie w podziale na płeć dla roku 2018 (agregacja po sekcji PKD i wieku).

**Widok PLATNICYSAM_SEKCJA2017**

Widok na tabeli PLATNICYSAM. Suma liczby płatników sam za siebie w podziale na sekcje dla roku 2017 (agregacja po płci i wieku).

**Widok PLATNICYSAM_SEKCJA2018**

Widok na tabeli PLATNICYSAM. Suma liczby płatników sam za siebie w podziale na sekcje dla roku 2018 (agregacja po płci i wieku).

**Widok PLATNICYSAM_SUMA2018**

Widok na tabeli PLATNICYSAM. Liczba płatników sam za siebie w gminie w roku 2018. Zawiera przybliżone wartości dla gmin o podwójnych nazwach, oszacowane na podstawie danych podatkowych (tabela pomocnicza zus.platnicy_podwojnegminy).

**Widok PLATNICYSAM_WIEK2017**

Widok na tabeli PLATNICYSAM. Suma liczby płatników sam za siebie w podziale na wiek dla roku 2017 (agregacja po sekcji PKD i płci)

**Widok PLATNICYSAM_WIEK2018**

Widok na tabeli PLATNICYSAM. Suma liczby płatników sam za siebie w podziale na wiek dla roku 2018 (agregacja po sekcji PKD i płci).

**Widok PODSTAWY_PLEC2017**

Widok na tabeli PODSTAWY. Średnie podstawy naliczenia składek emerytalnych w podziale na płeć dla roku 2017 (agregacja po sekcji PKD, wielkości płatnika i wieku ubezpieczonych).

**Widok PODSTAWY_PLEC2018**

Widok na tabeli PODSTAWY. Średnie podstawy naliczenia składek emerytalnych w podziale na płeć dla roku 2018 (agregacja po sekcji PKD, wielkości płatnika i wieku ubezpieczonych).

**Widok PODSTAWY_SEKCJA2017**

Widok na tabeli PODSTAWY. Średnie podstawy naliczenia składek emerytalnych w podziale na sekcje dla roku 2017 (agregacja po wielkości płatnika, płci i wieku ubezpieczonych).

**Widok PODSTAWY_SEKCJA2018**

Widok na tabeli PODSTAWY. Średnie podstawy naliczenia składek emerytalnych w podziale na sekcje dla roku 2018 (agregacja po wielkości płatnika, płci i wieku ubezpieczonych).

**Widok PODSTAWY_WIEK2017**

Widok na tabeli PODSTAWY. Średnie podstawy naliczenia składek emerytalnych w podziale na wiek dla roku 2017 (agregacja po sekcji PKD, wielkości płatnika i płci ubezpieczonych).

**Widok PODSTAWY_WIEK2018**

Widok na tabeli PODSTAWY. Średnie podstawy naliczenia składek emerytalnych w podziale na wiek dla roku 2018 (agregacja po sekcji PKD, wielkości płatnika i płci ubezpieczonych).

**Widok PODSTAWY_WIELKOSCPL2017**

Widok na tabeli PODSTAWY. Średnie podstawy naliczenia składek emerytalnych w podziale na wielkość płatnika dla roku 2017 (agregacja po sekcji PKD, wieku i płci ubezpieczonych).

**Widok PODSTAWY_WIELKOSCPL2018**

Widok na tabeli PODSTAWY. Średnie podstawy naliczenia składek emerytalnych w podziale na wielkość płatnika dla roku 2018 (agregacja po sekcji PKD, wieku i płci ubezpieczonych).

**Widok UBEZPIECZENI_SEKCJA2017**

Widok na tabeli UBEZPIECZENI. Suma liczby ubezpieczonych w podziale na sekcje PKD dla roku 2017 (agregacja po wielkości płatnika, wieku i płci).

**Widok UBEZPIECZENI_SEKCJA2018**

Widok na tabeli UBEZPIECZENI. Suma liczby ubezpieczonych w podziale na sekcje PKD dla roku 2018 (agregacja po wielkości płatnika, wieku i płci).

**Widok UBEZPIECZENI_WIEK2017**

Widok na tabeli UBEZPIECZENI. Suma liczby ubezpieczonych w podziale na wiek dla roku 2017 (agregacja po sekcji, wielkości płatnika i płci).

**Widok UBEZPIECZENI_WIEK2018**

Widok na tabeli UBEZPIECZENI. Suma liczby ubezpieczonych w podziale na wiek dla roku 2018 (agregacja po sekcji, wielkości płatnika i płci).

**Widok UBEZPIECZENI_WIELKOSCPL2017**

Widok na tabeli UBEZPIECZENI. Suma liczby ubezpieczonych w podziale na wielkość płatnika dla roku 2017 (agregacja po sekcji, wieku i płci).

**Widok UBEZPIECZENI_WIELKOSCPL2018**

Widok na tabeli UBEZPIECZENI. Suma liczby ubezpieczonych w podziale na wielkość płatnika dla roku 2018 (agregacja po sekcji, wieku i płci).

# Rozdział 2

Procedura importu danych udostępnionych przez Zakład Ubezpieczeń Społecznych i Krajową Administrację Skarbową oraz pobranych z Banku Danych Lokalnych do tabel docelowych. Spisana na potrzeby ewentualnej przyszłej aktualizacji zasobu o dane zebrane w kolejnych latach realizacji projektu.

## 2.1 Dane wejściowe

Dane wejściowe zostały przez ZUS i KAS udostępnione w postaci plików w formacie xls (xslx) o specyficznej strukturze, GUS publikuje dane w BDL w formacie xslx oraz csv, można wybrać format przy pobieraniu danych.

Informacje zagregowane są do gminy lub powiatu (niektóre zakresy danych BDL). Pojedynczy rekord tabeli źródłowej najczęściej dotyczy pojedynczej jednostki administracyjnej, ale np. w danych dot. liczby ubezpieczonych w gminie jeden rekord dotyczy liczby ubezpieczonych w danej gminie w przedsiębiorstwach danej wielkości. Należy upewnić się, że tabela zawiera kolumnę lub kombinację kolumn, które można potraktować jako jednoznaczny unikalny identyfikator. Może to być nazwa gminy (z oznaczeniem jej typu – uwaga na gminy o nieunikalnych nazwach) lub inny identyfikator. Jeśli nie występuje taki identyfikator rekordu, np. nazwa gminy występuje w rekordzie n i z kontekstu wynika, że rekordy n+1, n+2 itd. dotyczą tej samej gminy, ale po zmianie kolejności rekordów przyporządkowanie nie będzie możliwe, należy taki identyfikator dodać w kolejnej kolumnie.

Plik należy przygotować do importu do bazy danych, tzn.:

- Zmodyfikować nagłówki kolumn, tak aby spełniały warunki narzucone przez składnię języka SQL: nie zawierały spacji, znaków specjalnych, znaków interpunkcyjnych innych niż „_”, liter diakrytycznych, nie rozpoczynały się od liczby.
- Nagłówki kolumn muszą być krótkie, czytelne, unikalne w skali całej tabeli. Należy unikać kombinacji liter, które stanowią elementy składni języka SQL, takich jak „as”, „do”, „or” itp.
- Usunąć wszelkie opisy i wyjaśnienia z początkowych wierszy ponad nagłówkami kolumn.
- Oprogramowanie wykorzystywane do importu danych z xsl/csv do bazy danych może narzucać ograniczenie liczby kolumn w tabeli bazodanowej. Jeśli tak jest,

a tabela xsl/csv zawiera więcej kolumn (w szczególności dotyczy danych przekazywanych z ZUS) należy to uwzględnić, np. podzielić tabele źródłową na kilka tabel o mniejszej liczbie kolumn, oczywiście z zachowaniem we wszystkich tabelach kolumny zawierającej identyfikator rekordu (nazwa gminy w danych ZUS, identyfikator izby skarbowej w danych podatkowych, identyfikator gminy w danych GUS).

- Należy upewnić się, że w tabeli nie występują wartości błędne i przypadkowe, np. litery, spacje, myślniki w komórkach, w których powinny znajdować się tylko dane typu liczbowego.
- Warto usunąć kolumny i wiersze puste znajdujące się w środku tabeli, oraz kolumny i wiersze opisane jako „brak danych", „puste" itp.

## 2.2 Tabele robocze

Pliki przygotowane w sposób opisany wyżej należy zaimportować do bazy danych, do tabel roboczych, utworzonych w schemacie bazodanowym public lub w schemacie w którym znajdują się tabele docelowe (dla danych z ZUS, KAS, BDL odpowiednio schematy zus, podatki, bdl). Można wykorzystać wbudowaną funkcjonalność aplikacji wykorzystywanej do zarządzania bazą danych, oprogramowanie zewnętrzne posiadające funkcjonalność importu lub inny sposób zasilenia bazy danych. W zasileniu inicjalnym wykorzystano m.in. skrypty języka Phyton, realizujące tworzenie tabeli roboczej i import danych. Przykładowy skrypt importXls2Db.py stanowi załącznik do niniejszego dokumentu. Należy wykonać go w interpreterze języka Phyton, wykorzystano PhyCharm.

Po zasileniu tabel roboczych należy upewnić się, czy proces przebiegł prawidłowo; sprawdzić typy danych, liczbę kolumn, liczbę rekordów itp.

Tabele robocze wykorzystane w zasileniu inicjalnym po zakończeniu procesu importu uznano za zbędne i usunięto, aby ułatwić efektywne korzystnie ze schematu public.

## 2.3 Zapytania języka SQL i import do struktur docelowych

Gdy zasilono już tabele robocze należy dołączyć do danych identyfikator teryt (jeśli go nie posiadają), który stanowi klucz łączenia danych społeczno-gospodarczych z danymi przestrzennymi oraz przetworzyć dane z tabel roboczych do struktury tabel docelowych. Należy wykorzystać polecenia języka SQL: Select z narzuconymi warunkami, join. Następnie należy wykonać import wyniku zapytania do tabeli docelowej, przy wykorzystaniu konstrukcji: *INSERT INTO xxx () (SELECT xxx);*

## Dane KAS

Dane KAS prawdopodobnie nie wymagają skomplikowanych ingerencji w strukturę. W imporcie inicjalnym wykonano tylko rozdzielenie rekordu dla danej gminy na trzy, zawierające dane dla roku 2016, 2017 i 2018 oraz dołączenie identyfikatora teryt jak w poniższym przykładzie, zawierającym wybór informacji o deklaracji PIT28A w roku 2016:

insert into podatki.pit28a (teryt, rok, kod, gmina, r_firmy, liczba, poz10, poz11, poz12, poz13, poz8, poz9)

```
(select
        case
                when pit.gmina='Razem' then '04'
                else te.teryt
        end as teryt,
        '2016' as rok,
        pit.gmina_id as kod,
        pit.gmina as gmina,
        pit.r_firmy,
        pit.liczba_16 as liczba,
        pit.p10_16,
        pit.p11_16,
        pit.p12_16,
        pit.p13_16,
        pit.p8_16,
        pit.p9_16
                from public."PodatkiPIT28A_E" as pit
        left outer join
                public.teryt_kod as te
        on te.pkod=pit.gmina_id
where pit.liczba_16 is not null);
```

## Dane BDL

Dane BDL należało rozdzielić na podstawie przyjętych wyróżnień. Skrypt:

```
insert into bdl.zatrudnienie (teryt, kod, nazwa, rok, osob, procent, dzial)
(select t.teryt, z.* from
                public.teryt_kod t
        join
                        ((select
                        kod,
                        nazwa,
                        '2018' as rok,
                        ogolem_os as osob,
```

```
        null as procent,
        'ogolem' as dzial
        from public."BDLZatrudnienie2018_E")
union
        (select
        kod,
        nazwa,
        '2018' as rok,
        rolnictwo_os as osob,
        rolnictwo_proc as procent,
        'rolnictwo' as dzial
        from public."BDLZatrudnienie2018_E")
union
        (select
        kod,
        nazwa,
        '2018' as rok,
        przemysl_os as osob,
        przemysl_proc as procent,
        'przemysl' as dzial
        from public."BDLZatrudnienie2018_E")
union
        (select
        kod,
        nazwa,
        '2018' as rok,
        handel_os as osob,
        handel_proc as procent,
        'handel' as dzial
        from public."BDLZatrudnienie2018_E")
union
        (select
        kod,
        nazwa,
        '2018' as rok,
        finanse_os as osob,
        finanse_proc as procent,
        'finanse' as dzial
        from public."BDLZatrudnienie2018_E")
union
        (select
        kod,
        nazwa,
        '2018' as rok,
        pozostale_os as osob,
```

```
                pozostale_proc as procent,
                'pozostale' as dzial
                from public."BDLZatrudnienie2018_E")) as z
        on t.bkod=z.kod);
```

wybiera dane o zatrudnieniu w roku 2018 w podziale na działy gospodarki oraz przypisuje gminom identyfikator teryt.

### Dane ZUS

Dane ZUS przekazane zostały w postaci tabel o bardzo wielu kolumnach, pojedynczy wiersz zawiera dane dla całej gminy lub podmiotów danej wielkości w tej gminie, zaś kolejne komórki zawierają wartości w podziale na atrybuty: sekcja PKD, wielkość płatnika, wiek i płeć zatrudnionego. Dla pojedynczej gminy występuje nawet do 2000 wartości (dla bardzo złożonej tabeli Podstawy emerytalne ubezpieczonych), które w tabeli docelowej zapisać należy jako kolejne rekordy, opisując je kombinacją wartości wymienionych powyżej atrybutów.

Przykładowe zapytanie realizujące taką selekcje dla średniej podstawy emerytalnej mężczyzn poniżej 24 roku życia w przedsiębiorstwach zaklasyfikowanych w sekcji B dla roku 2017:

```
select
        case
                when gmina in ('Srednia') then '04'
                else te.teryt
        end as teryt,
        '2017' as rok,
        po.gmina,
        'B' as sekcja,
        case
                when trim(po.kategoria)=trim(po.gmina) then 'Ogolem'
                else po.kategoria
        end as wielkoscpl,
        'od24' as wiek,
        'M' as plec,
        po.b1m as kwota
from
        (select * from public.teryt_kod where nazwa<>'Rogowo') as te
        right outer join
        (select * from public."ZUS201712PodstawyEmerytalne_E" where gmina <> 'Suma końcowa')
as po
        on trim(te.znazwa)=trim(po.gmina)
 where po.b1m is not null;
```

Zapytanie takie, oraz insert do tabeli docelowej, należy wykonać dla wszystkich kombinacji wartości atrybutów sekcja, wiek i płeć, czyli około 400 razy. Jest to czasochłonne, uciążliwe i niesie duże ryzyko błędu – pominięcia lub wykonania wielokrotnie którejś sekwencji.

Zastosowano rozwiązanie półautomatyczne, z wykorzystaniem interpretera języka Phyton. Skrypt generuje i uruchamia sekwencję zapytań na podstawie dostarczonej listy nazw kolumn. Przykładowy skrypt ubezpieczeni2017.py stanowi załącznik do niniejszego dokumentu.

## 2.4 Przewidywane trudności:

1. W województwie występują gminy o nieunikalnych nazwach, co może powodować problem z jednoznacznym przypisaniem encji danych do gminy. W początkowej fazie projektu wystąpiły takie problemy.

   W danych BDL dane zawsze opisane są kombinacją nazwa + identyfikator jednostki statystycznej.

   W danych podatkowych dane opisane są nazwą gminy i identyfikatorem izby skarbowej. Przypisanie tego identyfikatora do identyfikatora TERYT zapisano w tabeli public.teryt_kod.

   W danych ZUS dane opisane są nazwą gminy. Na bieżącym etapie projektu nie było możliwe jednoznaczne przyporządkowanie tych rekordów, przekazano po jednym rekordzie dla danej nazwy gminy, więc prawdopodobnie zostały nadmiernie zagregowane. Należy zwrócić się do ZUS z prośbą o udostępnienie dodatkowego identyfikatora: identyfikatora teryt, a jeśli nie będzie to możliwe typu gminy i nazwy powiatu.

   Lista gmin o nieunikalnych nazwach, stan na koniec 2019:

| teryt gminy | nazwa gminy | rodzaj gminy | teryt powiatu | nazwa powiatu |
|---|---|---|---|---|
| 0401011 | Aleksandrów Kujawski | miejska | 0401 | aleksandrowski |
| 0401042 | Aleksandrów Kujawski | wiejska | 0401 | aleksandrowski |
| 0402032 | Brodnica | wiejska | 0402 | brodnicki |
| 0402011 | Brodnica | miejska | 0402 | brodnicki |
| 0404011 | Chełmno | miejska | 0404 | chełmiński |
| 0404022 | Chełmno | wiejska | 0404 | chełmiński |
| 0415011 | Chełmża | miejska | 0415 | toruński |
| 0415022 | Chełmża | wiejska | 0415 | toruński |
| 0405011 | Golub-Dobrzyń | miejska | 0405 | golubsko-dobrzyński |
| 0405032 | Golub-Dobrzyń | wiejska | 0405 | golubsko-dobrzyński |
| 0407011 | Inowrocław | miejska | 0407 | inowrocławski |
| 0407042 | Inowrocław | wiejska | 0407 | inowrocławski |
| 0418092 | Kowal | wiejska | 0418 | włocławski |
| 0418011 | Kowal | miejska | 0418 | włocławski |
| 0408062 | Lipno | wiejska | 0408 | lipnowski |
| 0408011 | Lipno | miejska | 0408 | lipnowski |
| 0411011 | Radziejów | miejska | 0411 | radziejowski |
| 0411062 | Radziejów | wiejska | 0411 | radziejowski |
| 0419052 | Rogowo | wiejska | 0419 | żniński |
| 0412032 | Rogowo | wiejska | 0412 | rypiński |
| 0412011 | Rypin | miejska | 0412 | rypiński |
| 0412042 | Rypin | wiejska | 0412 | rypiński |
| 0418132 | Włocławek | wiejska | 0418 | włocławski |
| 0464011 | Włocławek | miejska | 0464 | Włocławek |

2. Przy wykorzystywaniu zewnętrznych aplikacji do importu danych z tabel XLS do bazy danych, które automatycznie tworzą w BD tabele docelowe i automatycznie nadają typy danych kolumnom należy nadzorować poprawność tworzonej struktury. Zdarza się, że oprogramowanie błędnie rozpoznaje dane typu integer (liczby całkowite) jako typ real/numeric (liczby rzeczywiste, zmiennoprzecinkowe).