Riešitelia:

- Sabína Gregušová (xgregu02)
- Peter Hamran (xhamra00)
- Adrián Tulušák (xtulus00)

1 Úvod

2 Príprava dát

Tento projekt sa zameriava na situáciu v Českej republike (dotazy A, B) a na situáciu v Európskej únii (dotaz C) od vypuknutia pandémie. Keďže sa jedná o stále aktuálne téma, existuje k nemu pomerne veľké množstvo dát. Podarilo sa nám získať dátové sady, ktoré sú dostačujúce pre zodpovedanie všetkých 3 dotazov. Jedná sa o:

- pre A: COVID-19: Přehled osob s prokázanou nákazou dle hlášení krajských hygienických stanic (v2)
- pre B: COVID-19: Přehled epidemiologické situace dle hlášení krajských hygienických stanic podle okresu
- pre C:

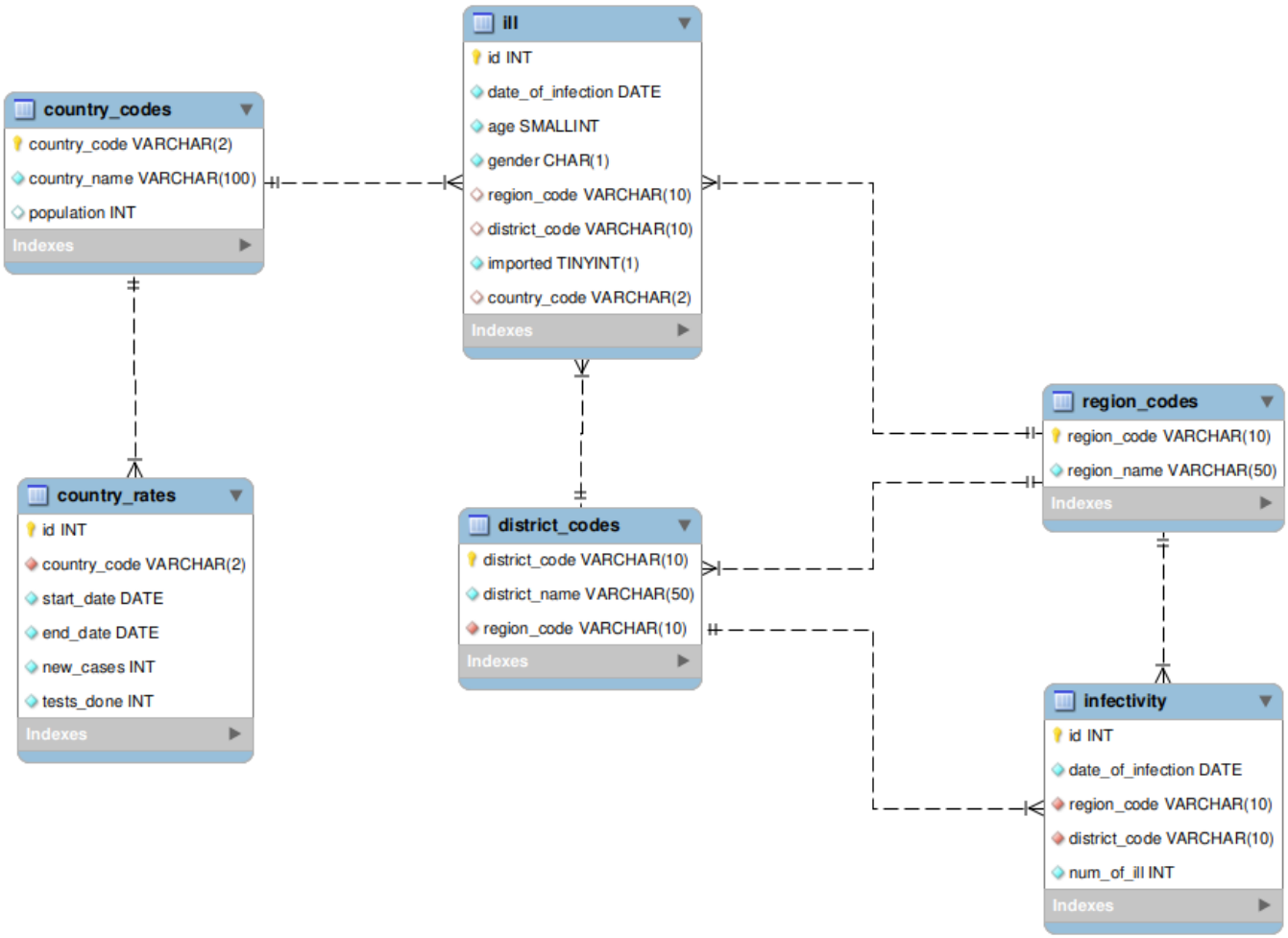
Tieto dátové sady obsahujú najpodstatnejšie informácie, no pre ich lepšiu prezentáciu boli vytvorené a pridané 2 dátové sady. Dotazy by bolo možné zodpovedať aj bez nich, no pridajú na finálnej zrozumiteľnosti. Jedná sa o:

- `countries.json` obsahuje unikátne skratky krajín a ich český ekvivalent
- `nuts_lau.json` obsahuje kódy NUTS (kraj) a LAU (okres) a ich český ekvivalent

Všetky dátové sady sú importované pomocou príkazu `mongoimport`. Dátové sady A a B potrebujú formátovanie, aby sme naozaj pracovali s podstatnými dátami. Celý tento proces je automatizovaný pomocou skriptu `prepData.sh`, ktorý:

- Zmaže a vytvorí novú (čistú) databázu s názvom ”corona”
- Postupne stiahne, a ak je to potrebné aj naformátuje, dátové sady vo formáte JSON
- Naimportuje každú dátovú sadu ako kolekciu; kolekcie sú pomenované podľa hlavnej dátovej sady (A, B, C) a novo pridané dátové sady sú pomenované ako countries a nuts_lau

Samotnú úpravu, prípravu a konvertovanie dát vykonáva skript `convertData.py`, ktorý sa spúšťa s Python verziou 3. Tento skript bol navrhnutý tak, aby bol ľahko modifikovateľný a rozšíriteľný a preto využíva objektové orientáciu. Na začiatku sa pripojí k NoSQL aj MySQL databázi. Následne vytvorí relačnú databázu v jazyku SQL - kód vytvorenia relačnej databáze sa nachádza v samostatnom súbore s názvom `Corona.sql`, ktorý je preparsovaný a vykonajú sa jednotlivé príkazy vrámci skriptu `convertData.py`. Toto riešenie zaisťuje, že samotný skript obsahuje minimum „natvrdo“ napísaných príkazov pre vytvorenie relačnej databáze.
Samotná relačná databáza obsahuje dokopy 6 tabuliek - 3 hlavné tabuľky pre zodpovedanie dotazov a 3 pomocné. Bolo by možné použiť iba 3 hlavné tabuľky, ale mnohé dáta by boli redundantné a neviedlo by to k vhodnému návrhu, preto boli vytvorené pomocné tabuľky. Slová ako `district` a `region` sú často zameniteľné, no v celom tomto projekte je slovo `district` považované za okres a `region` za kraj.



Obrázek 1: Schéma relačnej databázy