
# UPA - Projekt 1. část

Bc. Aleš Tetur (xtetur01),
Bc. Samuel Stuchlý (xstuch06),
Bc. Martin Litwora (xlitwo00)

Brno, 3. 11. 2021

# 1. Zadání

Našim cílem je analyzovat zdroje dat na námi zvolené téma. Dále máme analyzovat strukturu datových sad a návrh jejich uložení do vhodně zvolené databáze. Součástí řešení je i skript, který datové sady zpracuje a uloží do databáze.

# 2. Výběr tématu

Ze všeho nejdříve jsme prozkoumali všechny varianty zadání. Rozhodli jsme se pro variantu 02: Zdravotnictví v ČR. Dále jsme se seznámili s otázkami týkající se tohoto tématu na které budeme odpovídat v druhé části projektu.

# 3. Analýza a výběr technologie

Nejdříve jsme analyzovali možné zdroje dat, týkajících se našeho tématu. Prvně jsme zpracovali zdroj Národní registr poskytovatelů zdravotních služeb.

Tento zdroje jsme prozkoumali a analyzovali jsme formát datových sad. Data jsou ve formátu CSV a obsahují několik informací. Většinou se jedná a atomická data a celá sada je nestrukturalizovaná. Datová sada obsahuje tyto informace:

- ZdravotnickeZarizeniId int
- PCZ int
- PCDP int
- NazevCely varchar
- ZdravotnickeZarizeniKod bigint
- DruhZarizeniKod int
- DruhZarizeni varchar
- DruhZarizeniSekundarni varchar
- Obec varchar
- Psc int
- Ulice varchar
- CisloDomovniOrientacni varchar
- Kraj varchar
- KrajKod varchar
- Okres varchar
- OkresKod varchar
- SpravniObvod varchar
- PoskytovatelTelefon int
- PoskytovatelFax int
- DatumZahajeniCinnosti date
- IdentifikatorDatoveSchranky varchar
- PoskytovatelEmail varchar
- PoskytovatelWeb varchar
- DruhPoskytovatele varchar
- PoskytovatelNazev varchar
- Ico int
- TypOsoby varchar
- PravniFormaKod varchar
- KrajKodSidlo varchar
- KrajSidlo varchar
- OkresKodSidlo varchar
- OkresSidlo varchar
- PscSidlo int
- ObecSidlo varchar
- UliceSidlo varchar
- CisloDomovniOrientacniSidlo varchar
- OborPece varchar
- FormaPece varchar
- DruhPece varchar
- OdbornyZastupce varchar
- GPS varchar
- LastModified date

Některé informace jsou obsaženy pouze v některých datových sadách. Jedná se konkrétně o informace ZdravotnickeZarizeniKod, DruhZarizeniKod, DruhZarizeni, DruhZarizeniSekundarni. Níže je popsáno jak jsem tuto zkutečnost implementovali.

Dále jsme zpracovali data z druhého zdroje Data Českého statistického úřadu o obyvatelstvu ČR.

Tento zdroje jsme také prozkoumali a analyzovali jsme formát datových sad. Data jsou ve formátu CSV a obsahují několik informací. Některé informace nejsou známy. Jedná se o atomická data a celá sada je nestrukturalizovaná. Datová sada obsahuje tyto informace:

- idhod varchar
- hodnota int
- stapro_kod varchar
- pohlavi_cis varchar
- pohlavi_kod varchar
- vek_cis varchar
- vek_kod varchar
- vuzemi_cis varchar
- vuzemi_kod varchar
- casref_do date
- pohlavi_txt varchar
- vek_txt varchar
- vuzemi_txt varcha

Na základě analýzy jsme pro uložení dat zvolili technologii MongoDB. Do databáze této technologie se vkládají data mimo jiné ve formátu JSON nebo CSV. Bylo tedy zapotřebí navrhnout strukturu dat, která budeme ukládat. Oproti původní struktuře dat jsme sloučili informace, které spolu dávají komplexní informaci nebo spolu úzce souvisejí. Pro tuto úpravu jsme se rozhodli z důvodu jednoduššího vyhledávání těchto komplexních informací. Dále jsme se rozhodli nahrát data z různých zdrojů do vlastní kolekcí databáze a tím je od sebe oddělit. Struktura námi ukládaných dat je tedy:

- ZdravotnickeZarizeniId je “_id”
- PCZ
- PCDP
- NazevCely
- ZdravotnickeZarizeniKod
- DruhZarizeni
    - DruhZarizeniKod
    - DruhZarizeni
    - DruhZarizeniSekundarni
        - Může obsahovat více druhů
- AdresaZarizeni
    - Obec
    - Psc
    - Ulice
    - CisloDomovniOrientacni
    - Kraj
    - KrajKod
    - Okres
    - OkresKod
    - SpravniObvod
    - Poskytovatel
    - PoskytovatelTelefon
    - PoskytovatelFax
    - DatumZahajeniCinnosti
    - IdentifikatorDatoveSchranky
    - PoskytovatelEmail
    - PoskytovatelWeb
    - DruhPoskytovatele
    - PoskytovatelNazev
    - Ico
    - TypOsoby
    - PravniFormaKod
- Sidlo
    - KrajKodSidlo
    - KrajSidlo
    - OkresKodSidlo
    - OkresSidlo

    - PscSidlo
    - ObecSidlo
    - UliceSidlo
    - CisloDomovniOrientacniSidlo
- Pece
    - OborPece
    - FormaPece
    - DruhPece
- OdbornyZastupce
- GPS
- LastModified

Položky ZdravotnickeZarizeniKod, DruhZarizeniKod, DruhZarizeni, DruhZarizeniSekundarni jsou v datových sadách pouze od 1. 4. 2020. Proto v případě načítání datové sady z období před 1. 4. 2020 budou chybějící hodnoty vyplněné prázdnou hodnotou.

- Idhod je "_id"
- hodnota
- stapro_kod
- casref_do
- pohlavi
    - pohlavi_cis
    - pohlavi_kod
    - pohlavi_txt
- vek
    - vek_cis
    - vek_kod
    - vek_txt
- vuzemi
    - vuzemi_cis
    - vuzemi_kod
    - vuzemi_txt

Některé informace nejsou v datových sadách zadány a proto budou vyplněné prázdnou hodnotou. Společným atributem obou datových sad je území záznamu.

# 4. Implementace

Další částí řešení je vytvoření programu pro načtení, předzpracování a uložení dat do databáze. Pro implementaci tohoto programu jsme zvolili programovací jazyk Python, protože se v něm velmi jednoduše pracuje s CSV a JSON.

Samotný program nejdříve stáhne dvě datové sady ze zdroje a následně je začne zpracovávat. Pro jednodušší restrukturalizaci pracujeme uvnitř programu s daty ve formátu JSON. Datovou sadu Národní registr poskytovatelů zdravotních služeb jsme museli otevřít v kódování ISO-8859-2, jelikož funkce pro čtení souboru CSV implicitně pracuje s kódováním UTF-8. U druhého zdroje tento problém nebyl jelikož již je v UTF-8.

MongoDB databáze je vytvořena pomocí Docker obrazu, ve kterém je nainstalován MongoDB. Do databáze vkládáme záznamy postupně, tedy každý řádek z původní sady zvlášť. Pro úpravu struktury jsme si vytvořili vlastní funkce, které původní data restrukturalizují. Po nahrání dat, obsahuje databáze dvě kolekce. První kolekce *poskytovateleZP* obsahující data první sady a druhá kolekce *obyvatelstvo* obsahující data druhé sady.

GitHub repozitář projektu:
https://github.com/SamuelStuchly/UPA2021

# 5. Zprovoznění

Pro fungování je nutné mít nainstalován *Python3* a *Docker*.
Před spuštěním je nutné spustit

*make clean*

dále vytvořit virtuální prostředí pomocí příkazu

*make venv*

*source venv/bin/activate*

spustit skript příkazem

*make*

Podrobnější postup je v souboru README.md