UPA - projekt - 1. část

Zvolené téma

COVID-19

Řešitelé

- Jakub Frejlach - xfrejl00
- Tomáš Sasák - xsasak01
- Tomáš Venkrbec - xvenkr01

Zvolené dotazy a formulace vlastního dotazu

- **Skupina A** - v grafech zobrazte tempo změny počtů aktuálně nemocných (absolutní i procentuální přírůstek pozitivních případů a klouzavý průměr různých délek v různých časech)
- **Skupina B** - určete vliv epidemie COVID-19 na počet zemřelých v porovnání dle počtu nemocných, počtu hlášených úmrtí na nemoc COVID-19 a v porovnání s minulými lety
- **Vlastní dotaz** - zobrazte vývoj poměru vyléčených a zemřelých v různých časech

Stručná charakteristika zvolené datové sady

0.0.1 Datové sady

1. COVID-19: Celkový (kumulativní) počet osob s prokázanou nákazou dle krajských hygienických stanic včetně laboratoří, počet vyléčených, počet úmrtí a provedených testů - tento soubor dat nám poskytne důležité informace k dotazům skupiny A, B i k vlastnímu dotazu.

   Položky datové sady:

   - `datum` - konkrétní datum, ke kterému se vážou následující položky
   - `kumulativni_pocet_nakazenych` - kumulativní počet nakažených v daný den
   - `kumulativni_pocet_vylecenych` - kumulativní počet vyléčených v daný den
   - `kumulativni_pocet_umrti` - kumulativní počet úmrtí v daný den
   - `kumulativni_pocet_testu` - kumulativní počet testů v daný den

2. Zemřelí podle týdnů a věkových skupin v České republice - tento soubor dat bude důležitý zejména pro zodpovězení dotazu skupiny B.

   Položky datové sady:

   - `idhod` - unikátní identifikátor údaje Veřejné databáze ČSÚ, využije se v případě dotazu ke konkrétnímu údaji
   - `hodnota` - zjištěná hodnota, v numerickém formátu
   - `stapro_kod` - kód statistické proměnné ze systému SMS UKAZ, v této DS pouze kód 5393 pro Počet zemřelých osob s trvalým nebo dlouhodobým pobytem
   - `vek_cis` - kód číselníku pro věkovou skupinu, využit číselník 7700, pokud není vyplněn, jedná se o úhrn za všechny věkové skupiny
   - `vek_kod` - kód položky číselníku pro věkovou skupinu

- `vuzemi_cis` - kód číselníku pro referenční území, číselník odpovídá typologii území, pro údaj za stát použit číselník 97
- `vuzemi_kod` - kód položky číselníku pro referenční území, pro údaj za Českou republiku kód 19
- `rok` - rok referenčního období ve formátu RRRR, rok pro referenční týden dle normy ISO
- `tyden` - pořadové číslo referenčního týdne, dle normy ISO
- `roktyden` - referenční rok a týden dle normy ISO, formát RRRR-Wxx
- `casref_od` - datum odpovídající prvnímu dni (pondělí) referenčního týdne, ve formátu RRRR-MM-DD
- `casref_do` - datum odpovídající poslednímu dni (neděle) referenčního týdne, ve formátu RRRR-MM-DD
- `vek_txt` - text pro věkovou skupinu

Způsob získání dat

Získání těchto datových sad bude provedeno pomocí HTTP dotazu a data ve formátu CSV nebo JSON budou zpracovány pomocí programu v jazyce Python 3 a uložené v NoSQL databázi.

První uvedený datový soubor má strukturu kumulativního počtu nakažených, vyléčených a zemřelých osob v jednotlivých dnech.

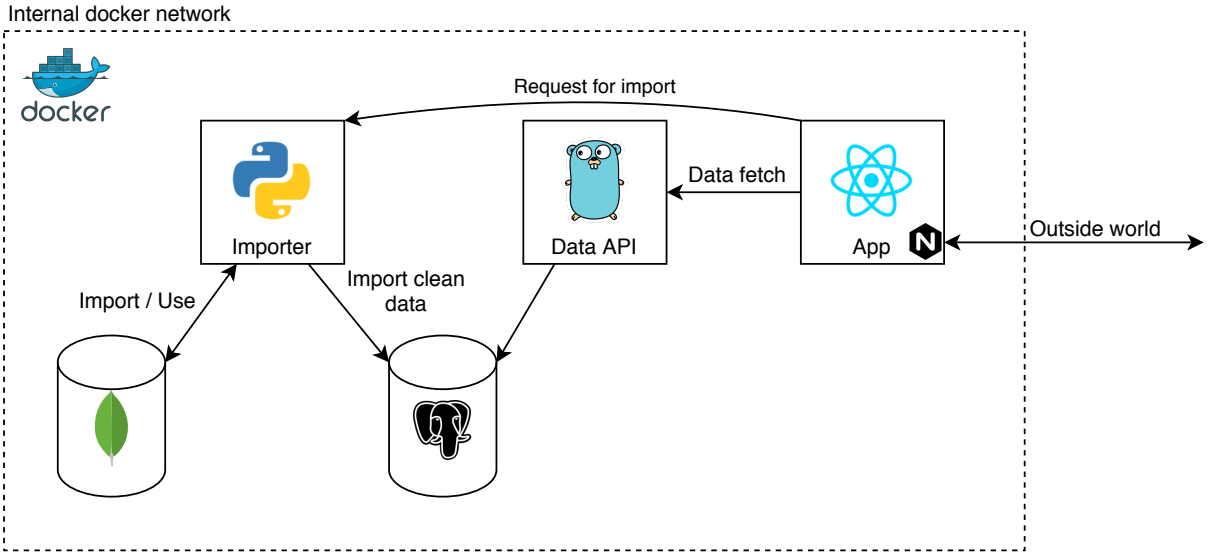
Druhý datový soubor má strukturu počtu úmrtí v jednotlivých týdnech rozdělený do věkových kategorií. K dispozici je i celkový počet zemřelých pro daný týden. Pracovat budeme s počty zemřelých a s obdobím. Datová sada zaznamenává úmrtí od roku 2011 až doposud.

Zvolený způsob uložení surových dat

Pro uložení získaných dokumentů pro jejich další zpracování nám poslouží NoSQL databáze MongoDB. Jedná se o nerelační databázi, kde jsou dokumenty reprezentované ve formátu JSON. Konkrétně jde o nerelační dokumentovou databázi. Tuto databázi jsme vybrali hlavně z důvodu, že velká část našich dat jsou časové údaje, na které je MongoDB vhodná. Dalším důvodem byla všeobecně velká rozšířenost MongoDB a tudíž i větší dostupných informací.

UPA - projekt - finální část

Architektura aplikace



Obrázek 1: Schéma architektury aplikace

Celá aplikace se skládá z pěti oddělených částí v jednotlivých Docker kontejnerech. Jednotlivé části spolu komunikují podle schématu na obrázku 1. Bližší popis jednotlivých částí a způsobu komunikace s okolím:

1. **MongoDB** (`upa-mongo`) - kontejner obsahující instanci nerelační dokumentové databáze MongoDB. Databáze komunikuje datově s `importerem` v obou směrech. Surová data jsou zde nejprve `importerem` uložena a posléze jsou odtud opět získána pro další zpracování.

2. **Postgres** (`upa-postgres`) - kontejner obsahující instanci relační databáze Postgres. `Importer` sem ukládá zpracovaná a očištěná data. Rovněž odtud datové `API` získává data.

3. **Importer** (`upa-importer`) - kontejner obsahující aplikaci napsanou v jazyce Python 3, která slouží jako prostředník pro režii při získávání, zpracování a ukládání dat. Jedná se vlastně o `AIOHTTP REST API`. Základní funkce `importeru` je získání dat z veřejně dostupných zdrojů popsaných v sekci Datové sady, uložení těchto dat v `MongoDB`, získání dat z `MongoDB` a jejich následné zpracování a očistění a konečně uložení těchto očištěných dat v `Postgres` databázi. Činnost `importeru` je vyvolána pomocí HTTP dotazu na jeden ze dvou endpointů s příslušnými parametry (viz `README.md`).

4. **Data API** (`upa-api`) - kontejner obsahující aplikaci napsanou v jazyce Golang s využitím frameworků GIN a GORM. Jde o datové API, které na základě příchozích HTTP dotazů se správnými parametry (viz `README.md`) získá pomocí SQL dotazů požadovaná data z `Postgres` databáze a ve formátu JSON je odešle v HTTP zprávě.

5. **App** (`upa-app`) - kontejner obsahující frontend celé aplikace napsaný v Javascriptovém frameworku React. Webová aplikace obsahuje 3 grafy, které slouží jako odpovědi na dotaz skupiny A a vlastní dotaz. Oba grafy mají nastavitelný časový interval pro zobrazovaná data a u jednoho grafu lze navíc nastavit krok klouzavého průměru. Data pro grafy jsou získány pomocí HTTP dotazů na datové API. Aplikace je rovněž schopna vynutit činnost `importeru` pomocí tlačítek v záhlaví stránky.

Pokud nechcete aplikaci spouštět lokálně a instalovat závislosti, je možné si jí vyzkoušet online na tomto linku: http://165.227.168.39/

Zpracování dat (data uložené a připravené v SQL databázi)

1. Data z první datové sady, tedy **Celkový (kumulativní) počet osob s prokázanou nákazou dle krajských hygienických stanic včetně laboratoří, počet vyléčených, počet úmrtí a provedených testů**, byla zpracována a očištěna tak, aby obsahovala nikoliv kumulativní počty ale konkrétní počty testů, úmrtí, atd. pro daný den.

   Schéma tabulky `covid19` v Postgres databázi:

   - `id` - PK tabulky
   - `date_` - datum konkrétního dne, ke kterému se zbývající data vážou
   - `infected` - počet nově nakažených v daný den
   - `cured` - počet vyléčených v daný den
   - `deaths` - počet úmrtí v souvislosti s onemocněním COVID-19 v daný den
   - `ts` - časový údaj o tom, kdy byl konkrétní řádek vložen do tabulky

2. Data z druhé datové sady, tedy **Zemřelí podle týdnů a věkových skupin v České republice**, byla zpracována a očištěna tak, aby bylo možné zbavit se nepotřebných položek a převést časové údaje do přívětivějších formátů.

   Schéma tabulky `deaths` v Postgres databázi:

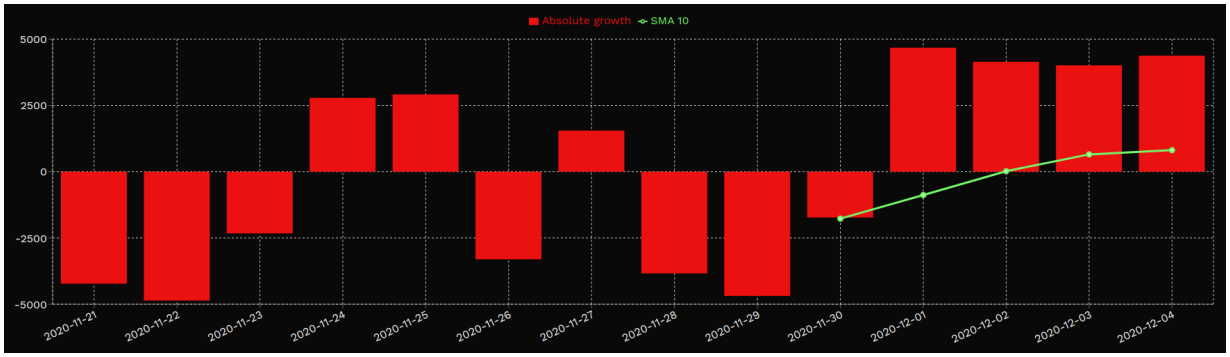
   - `id` - PK tabulky
   - `date_from` - datum od kterého se příslušný počet úmrtí počítá
   - `date_to` - datum do kterého se příslušný počet úmrtí počítá (vždy se jedná o rozmezí jednoho týdne)
   - `week` - pořadí týdne v roce
   - `deaths` - počet úmrtí
   - `age_from` - spodní hranice věkové kategorie pro danou hodnotu úmrtí
   - `age_to` - horní hranice věkové kategorie pro danou hodnotu úmrtí (hodnota `Infinity` je použita pro reprezentování horní hranice "85 let a více")

Výsledky

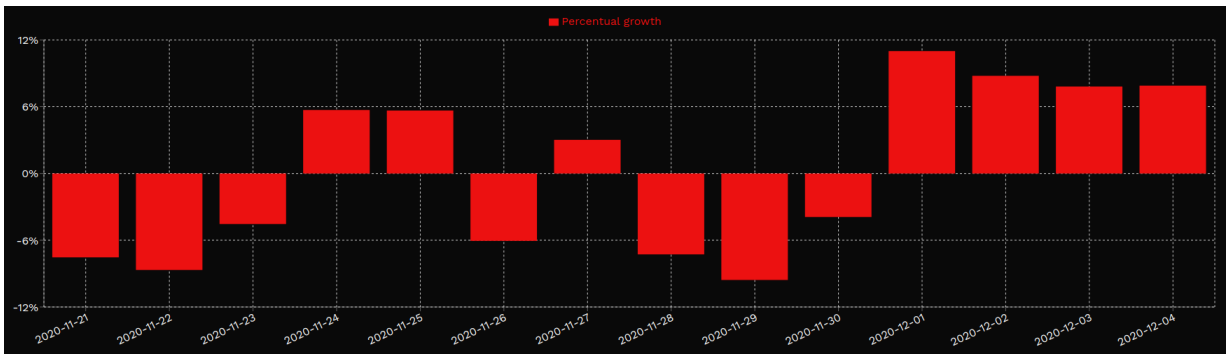
Dotaz skupiny A

V grafech zobrazte tempo změny počtů aktuálně nemocných (absolutní i procentuální přírůstek pozitivních případů a klouzavý průměr různých délek v různých časech)

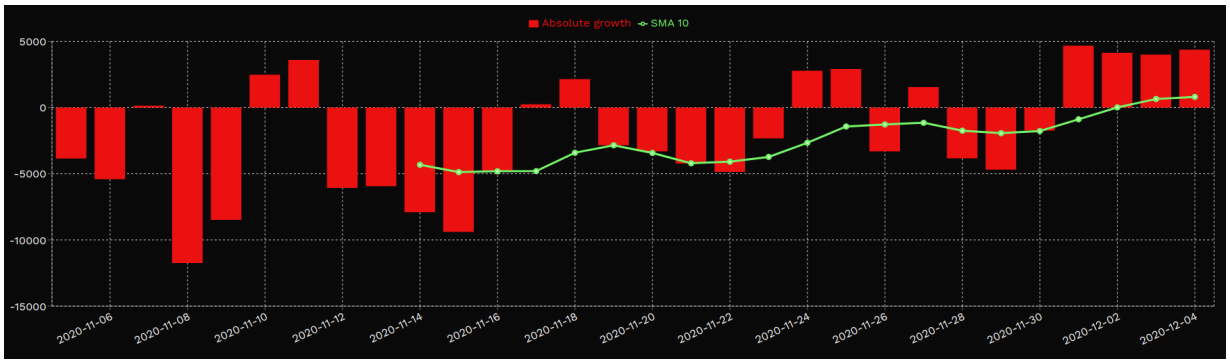
Pro ilustraci odpovědí na tento dotaz byly využity obrázky grafů získané z naší aplikace, konkrétně se jedná o tři dvojice grafů, které ukazují vždy po řadě absolutní přírůstek nemocných společně s klouzavým průměrem o velikosti kroku 10 a procentuální přírůstek nemocných. Jedná se o časové úseky posledních 14 dnů, posledního měsíce a od počátku pandemie COVID-19. Posledních 14 dní a poslední měsíc jsou brány vzhledem k datu vzniku tohoto dokumentu. Pro bližší analýzu dat, zobrazení konkrétních hodnot a možnost volby časového intervalu dat a kroku klouzavého průměru využijte prosím přiloženou aplikaci.



Obrázek 2: Absolutní přírůstek nemocných a klouzavý průměr s krokem 10 za posledních 14 dnů

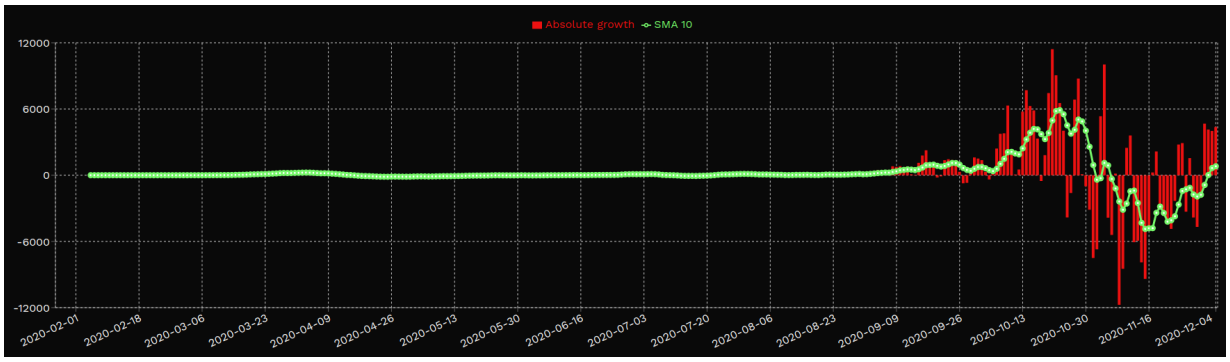


Obrázek 3: Procentuální přírůstek nemocných za posledních 14 dní

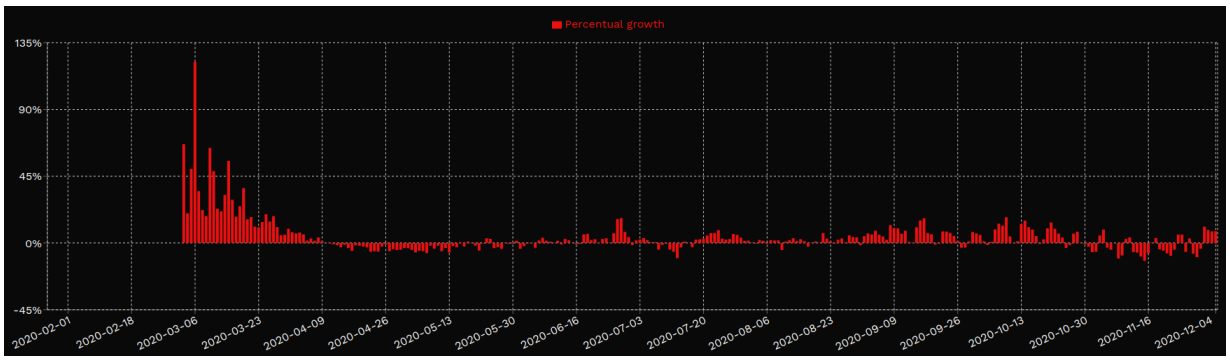


Obrázek 4: Absolutní přírůstek nemocných a klouzavý průměr s krokem 10 za poslední měsíc

Obrázek 5: Procentuální přírůstek nemocných za poslední měsíc



Obrázek 6: Absolutní přírůstek nemocných a klouzavý průměr s krokem 10 od počátku pandemie

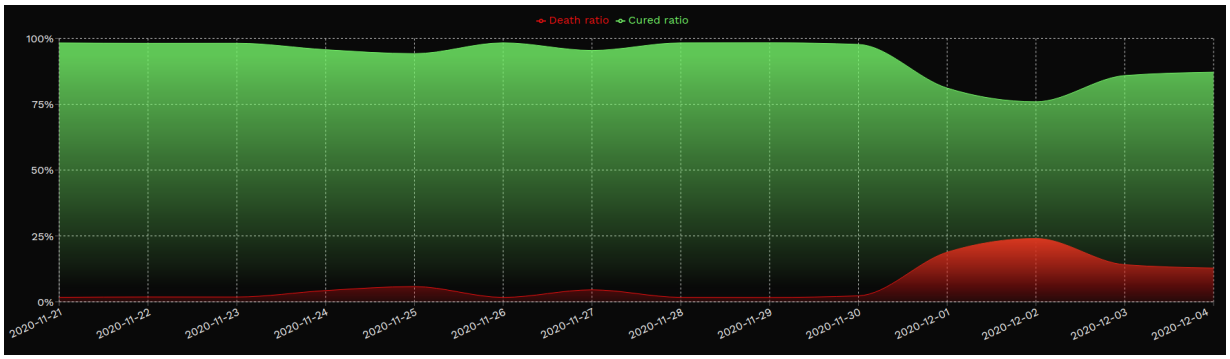


Obrázek 7: Procentuální přírůstek nemocných od počátku pandemie

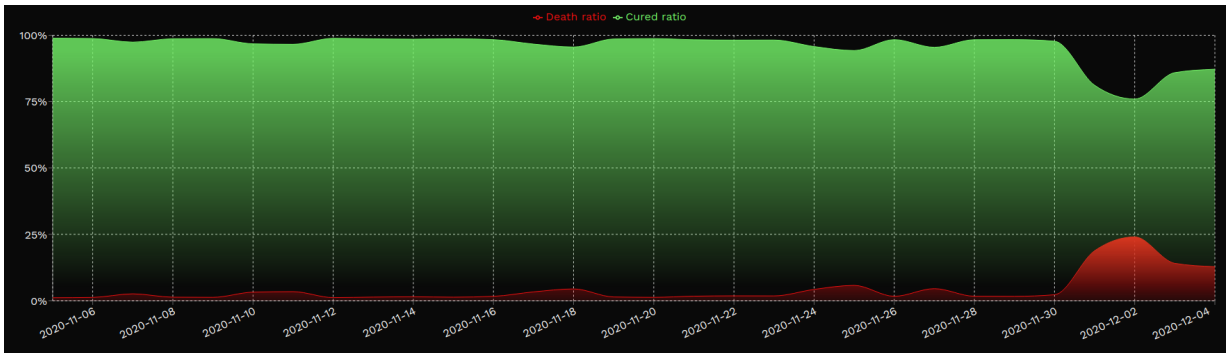
Vlastní dotaz

Zobrazte vývoj poměru vyléčených a zemřelých v různých časech.

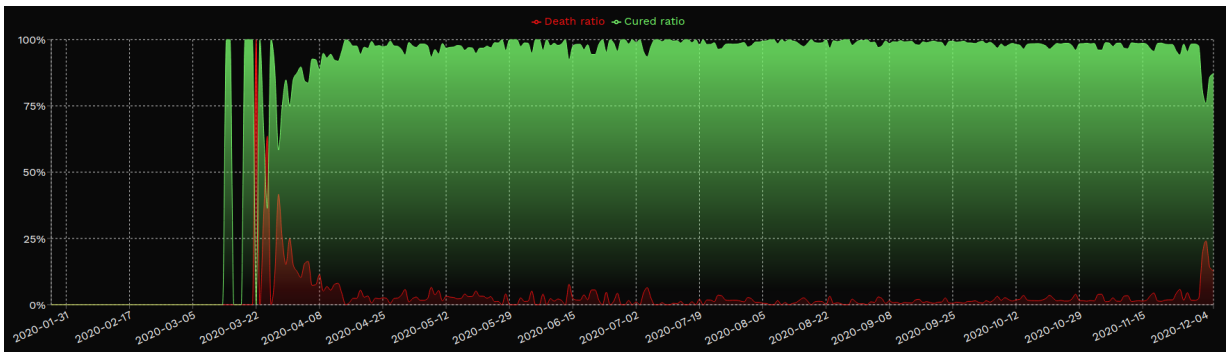
Pro ilustraci odpovědí na tento dotaz byly využity obrázky grafů získané z naší aplikace, konkrétně se jedná o tři grafy zobrazující poměr vyléčených a zemřelých v časových úsecích posledních 14 dnů, posledního měsíce a od počátku pandemie COVID-19. Pro bližší analýzu dat, zobrazení konkrétních hodnot a možnost volby časového intervalu dat využijte prosím přiloženou aplikaci.



Obrázek 8: Poměr počtu vyléčených a zemřelých za posledních 14 dní



Obrázek 9: Poměr počtu vyléčených a zemřelých za poslední měsíc



Obrázek 10: Poměr počtu vyléčených a zemřelých od počátku pandemie COVID-19