Zvolené téma: 4 - COVID-19
Řešitelé: Kateřina Fořtová (xforto00), Jakub Kolb (xkolbj00), Ondřej Pavela (xpavel34)

Zvolené dotazy a formulace vlastního dotazu:

- **Dotaz skupiny A** - vytvořte popisné charakteristiky pro alespoň 4 údaje (např. věk, pohlaví, okres, zdroj nákazy) z datové sady COVID-19: Přehled osob s prokázanou nákazou dle hlášení krajských hygienických stanic (využijte krabicové grafy, histogramy, atd.)
- **Dotaz skupiny 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 počtu testů a počtu vyléčených, dále zobrazte počty mrtvých dle krajů

Stručná charakteristika zvolené datové sady:

(Zde konkrétně popište jaké soubory budou představovat zdroj dat pro zvolené úlohy. Dále popište, jakým způsobem budou tato data získána a stručně charakterizujte strukturu souborů vybraných pro řešení projektu. Zaměřte se na části souborů, které jsou důležité pro zodpovězení zvolených dotazů.)

Dotaz skupiny A:

Pro dotaz skupiny A bude využito datové sady, kterou poskytuje Ministerstvo zdravotnictví ČR (https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/osoby.json). Tato datová sada je poskytována ve formátech *JSON* a *CSV*, přičemž v našem případě jsme zvolili formát *JSON*. Datový soubor se poté skládá z hlavičky a ze samotného seznamu osob s prokázanou nákazou. Hlavička obsahuje datum a čas poslední modifikace, které využijeme pro odlišení nových záznamů při aktualizacích naší NoSQL databáze.
Formát pro jednu konkrétní osobu s potvrzenou nákazou je následující:

```
{
	"datum": "2020-10-07",
	"vek": 55,
	"pohlavi": "Z",
	"kraj_nuts_kod": "CZ072",
	"okres_lau_kod": "CZ0721",
	"nakaza_v_zahranici": false,
	"nakaza_zeme_csu_kod": ""
}
```

Položka `"datum"` udává, kdy byla nákaza prokázána krajskou hygienickou stanicí.
Položky **`'kraj_nuts_kod'`** a **`'okres_lau_kod'`** udávají okres a jeho nadřízený kraj, v němž byla nákaza prokázána. Položka udávající kraj je zde redundantní v případě, že máme k dispozici informaci o územním dělení České republiky. Poslední dvě položky

`"nakaza_v_zahranici"` a `"nakaza_zeme_csu_kod"`, indikují, zda došlo k nákaze v zahraničí a pokud ano, tak v jaké zemi, dle kódu z číselníku ČSÚ.

Dotaz skupiny B:

Pro dotaz skupiny B je nutno využít více datových sad. První z nich je opět sada poskytovaná MZČR (https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/umrti.json), která je dostupná ve formátech *JSON* a *CSV*. Opět jsme zvolili formát *JSON*, který má stejnou strukturu jako sada z dotazu A.

Formát jedné datové položky je následující:

```
{
        "datum": "2020-03-24",
        "vek": 44,
        "pohlavi": "M",
        "kraj_nuts_kod": "CZ080",
        "okres_lau_kod": "CZ0802"
}
```

Položka `"datum"` udává, kdy bylo úmrtí zaznamenáno. Význam zbylých položek je obdobný jako v předchozím případě.

Dále bude nutné využít datovou sadu z jiného zdroje, která nám bude podávat informace o vývoji mortality v předešlých letech v ČR. Pro tuto skutečnost jsme využili datové sady poskytované Českým statistickým úřadem, která udává týdenní počty zemřelých pro každý týden v roce již od roku 2011 (https://www.czso.cz/documents/62353418/138258837/130185-20data092920.csv). Tyto data jsou zapisována do datové sady s určitým zpožděním (např. v říjnu 2020 máme poslední data dostupná počátkem srpna 2020). Vzhledem ke skutečnosti, že tato datová sada je dostupná pouze ve formátu *CSV*, bylo potřeba vynaložit více úsilí při předzpracování, abychom získali jednotlivé datové objekty ve slovníkovém formátu. Formát datové sady je následující:

```
idhod,hodnota,stapro_kod,vek_cis,vek_kod,vuzemi_cis,vuzemi_kod,rok,tyden,roktyden,casref_od,casref_do,vek_txt
832459719, 9, 5393, 7700, 400000610015000, 97, 19, 2011, 01, 2011-W01, 2011-01-03, 2011-01-09, 0-14
832460200, 55, 5393, 7700, 410015610040000, 97, 19, 2011, 01, 2011-W01, 2011-01-03, 2011-01-09, 15-39
832460678, 457, 5393, 7700, 410040610065000, 97, 19, 2011, 01, 2011-W01, 2011-01-03, 2011-01-09, 40-64
832461162, 1151, 5393, 7700, 410065610085000, 97, 19, 2011, 01, 2011-W01, 2011-01-03, 2011-01-09, 65-84
832461643, 577, 5393, 7700, 410085799999000, 97, 19, 2011, 01, 2011-W01, 2011-01-03, 2011-01-09, 85 a více
832462584, 2249, 5393, , , 9, 19, 2011, 01, 2011-W01, 2011-01-03, 2011-01-09, celkem
```

Stručný popis významu jednotlivých sloupců:

- **idhod:** ČSÚ identifikátor záznamu
- **hodnota:** počet zemřelých
- **stapro_kod:** kód statistické proměnné, zde pouze 5393
- **vek_cis:** číselník pro věkovou skupinu, zde pouze 7700
- **vek_kod:** kód položky číselníku pro věkovou skupinu
- **vuzemi_cis:** kód číselníku pro referenční území, pouze 93 (stát)
- **vuzemi_kod:** kód položky číselníku pro referenční území, pouze 19 (ČR)
- **rok:** rok referenčního období ve formátu RRRR
- **tyden:** pořadové číslo referenčního týdne (dle normy ISO)

- **roktyden:** referenční rok a týden ve formátu RRRR-Wxx (dle normy ISO)
- **casref_od:** datum pondělí referenčního týdne ve formátu RRRR-MM-DD
- **casref_do:** datum neděle referenčního týdne ve formátu RRRR-MM-DD
- **vek_txt:** text udávající věkovou skupinu, 0-14, 15-39, 40-64, 65-84, 85+

Hodnota v posledním řádku daného referenčního týdne vždy udává celkový úhrn úmrtí za všechny věkové skupiny.

Vlastní dotaz

Jako vlastní dotazy jsme si vybrali zobrazení vývoje testování a vyléčených případů na nemoc COVID-19. Dále jsme si vybrali zobrazení počtu mrtvých podle jednotlivých krajů. Pro zjištění počtu testů v jednotlivých dnech jsme využili datovou sadu opět od MZČR (https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/testy.json). Jedná se znovu o data v preferovaném formátu *JSON*, které pro náš projekt byly nejsnazší na zpracování. Níže je ukázka jednoho datového záznamu:


```
{
            "datum": "2020-01-29",
            "prirustkovy_pocet_testu": 5,
            "kumulativni_pocet_testu": 33
}
```


Pro náš graf jsme využili údaj přírustkového počtu testů za dané datum.
Pro informace o vyléčených jsme využili kumulativní statistiky opět dostupné ze stejného zdroje:
(https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/nakazeni-vyleceni-umrti-testy.json). Tato datová sada je opět ve formátu *JSON*. Následuje ukázka jednoho záznamu z této datové sady:


```
{
            "datum": "2020-07-08",
            "kumulativni_pocet_nakazenych": 12829,
            "kumulativni_pocet_vylecenych": 9771,
            "kumulativni_pocet_umrti": 353,
            "kumulativni_pocet_testu": 584946
 }
```


Položky `"datum"` a `"kumulativni_pocet_vylecenych"` budou sloužit jako údaje na základě kterých zobrazíme finální graf.

Pro poslední vlastní dotaz zobrazující počet úmrtí dle jednotlivých krajů jsme využili stejně jako u dotazu B datovou sadu poskytovanou MZČR (https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/umrti.json). Opět jsme využili informací o územním dělení ČR a na základě daného kraje máme tedy k dispozici všechny potřebné informace pro vykreslení grafu.

Získání a zpracování dat

Pro získání a zpracování datových sad jsme zvolili skriptovací jazyk *Python*, jelikož umožňuje jednoduchou práci se soubory formátu *JSON,* které se zde přímo mapují na vestavěné slovníky. Kromě toho existuje pro tento jazyk mnoho vyspělých knihoven pro jednoduché stahování souborů z internetu. Konkrétně jsme zvolili knihovnu *requests* (https://requests.readthedocs.io/en/master/), která podporuje vše, co jsme pro naše účely potřebovali.
Pro datové sady poskytované od MZČR zahrnovalo následné zpracování pouze vyřazení vybraných datových položek (viz dále) a vyfiltrování záznamů na základě data v případě aktualizace již existující databáze.
Při zpracování datové sady od ČSÚ ve formátu *CSV* bylo potřeba data výrazněji transformovat, aby následně odpovídala navrženému schématu naší NoSQL a relační databáze. Datový soubor je proto zpracovávan po jednotlivých týdnech jako celcích. Nejprve jsou vyřazeny nepotřebné sloupce: **idhod, stapro_kod, vek_cis, vek_kod, vuzemi_cis, vuzemi_kod, roktyden, vek_txt**. Následně je všech 6 záznamů za jeden konkrétní týden sloučeno do jednoho datového slovníku, který poté obsahuje počty úmrtí všech věkových skupin (celkový počet je možné dopočítat a není proto zahrnut).

Při práci s daty jsme upřednostňovali data ve formátu JSON díky snadnému stahování a získávání potřebných údajů pomocí skriptu napsaném v jazyce Python.

Zvolený způsob uložení uložení surových dat:

(Zde stručně charakterizujte NoSQL databázi, která bude využita pro uložení zvolených zdrojových dat.)

Pru uložení nestrukturovaných data jsme si zvolili dokumentovou NoSQL databázi MongoDB. Pro naše potřeby jsme poté vytvořili tři datové kolekce pro uložení datových sad od MZČR:

- **`rho_confirmed_cases`:** zde jsou uloženy veškeré osoby s potvrzenou nákazou. Nejprve byly vyřazeny nepotřebné datové položky (v našem případě pouze kód kraje) z jednotlivých záznamů osob reprezentovaných slovníky a výsledný seznam slovníků byl poté uložen do kolekce, což bylo velmi jednoduché, díky přímému mapování Python slovníku na MongoDB záznam.
- **`covid_deaths`:** kolekce uchovávající záznamy o úmrtích na COVID-19. Zpracování je obdobné jako v předchozím případě (opět byl vyřazen pouze kód kraje z jednotlivých záznamů)
- **`daily_statistics`:** kolekce sdružující více MZČR datových sad s charakterem časové řady. Konkrétně jsou zde uloženy záznamy z následujících zdrojů:
    - https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/testy.json
    - https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/nakaza.json
    - https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/nakazeni-vyleceni-umrti-testy.json

Záznamy bylo možné propojit na základě sdílené datové položky **'datum'**, která je v tomto případě jednoznačným identifikátorem. Pro každé datum byl vytvořen jeden záznam, který poté sdružoval vybrané datové položky ze všech sad, konkrétně:

- **'prirustkovy_pocet_testu'**
- **'prirustkovy_pocet_nakazenych'**
- **'kumulativni_pocet_vylecenych'**
- **'kumulativni_pocet_umrti'**

Je zřejmé, že tyto datové sady mají charakter časové řady, nicméně se nám nechtělo kombinovat více NoSQL databází pro různé části dat. Navíc je datová sada rozšiřována s frekvencí jednoho záznamu na jeden den, což výrazně limituje její velikost a využití specializované NoSQL databáze typu InfluxDB je proto dle nás v takovém případě prozatím neopodstatněné.

- **weekly_deaths:** zde je uložena datová sada od ČSÚ, jejíž způsob předzpracování byl popsán v předchozí sekci. Výsledné datové slovníky, jsou zde pouze uloženy do kolekce.

Vzhledem k tomu, že veškerá data o osobách jsou anonymní a neobsahují žádné unikátní identifikátory, nebylo příliš mnoho možností, na základě čeho jednotlivé datové sady propojit. Stále je ovšem možné jednoznačně propojit záznamy na základě časového rozmezí při tvorbě anonymních statistik, což je v našem případě dostačující.
Absence unikátních identifikátorů nicméně představuje problém při pozdější aktualizaci některých kolekcí. Konkrétně datové sady poskytované MZČR jsou průběžně aktualizovány v průběhu dne, což by samo o sobě ještě nebylo problémem, jelikož jsme při importu schopni ignorovat záznamy z aktuálního dne. Bohužel jsou v mnoha případech zpětně vkládány starší záznamy i s několikadenním zpožděním. To má za následek to, že nejsme při aktualizaci vlastní databáze schopni rozpoznat nově vložené záznamy s datem starším než je datum poslední aktualizace. Tím je do určité míry porušena konzistence a jediná možnost, jak tento problém řešit je opětovný import celé datové sady.