

# FAKULTA INFORMAČNÍCH TECHNOLOGIÍ VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Ukládání a příprava dat — projekt  
Zdravotnictví v ČR

# Úvod

Z nabízených témat bylo vybráno téma číslo 2: Zdravotnictví v ČR. Cílem projektu je seznámit se se zpracováním rozsáhlých/nestrukturovaných dat, seznámit se s přípravou těchto dat pro další využití (např. pro získávání znalostí z databází) a tvorbou popisných charakteristik pro zvolená data.

## 1 Návrh zpracování a uložení dat

Tato část projektu se zabývá seznámením se s vybraným tématem, analýzou dat z nabízených zdrojů, návrhu způsobu získání datových sad z daných zdrojů, výběru vhodné NoSQL databáze a uložení dat do zvolené databáze.

Byla zkoumána data ze zadaných zdrojů (Národní registr poskytovatelů zdravotních služeb a data Českého statistického úřadu o obyvatelstvu ČR). Na základě analýzy datových sad a jejich struktury byla pro jejich uložení zvolena NoSQL databáze MongoDB. Data nejsou tvořena časovými řadami, takže nevyhovuje InfluxDB, ani se nehodí uložení dat pomocí grafové databáze Neo4j.

### 1.1 Načtení dat ze zdrojů

Pro získání dat ze zadaných zdrojů byl vytvořen skript `download.py`. Tento skript také nahrává data do cloudové databáze.

Pro většinu zadaných dotazů jsou potřeba pouze aktuální data, ale pro druhý dotaz skupiny A je potřeba získat i historii poskytovatelů zdravotních služeb. Při práci z daty bylo zjištěno, že historická data jsou velice objemná a je problematické je uložit do použité cloudové databáze. Tento problém byl aktuálně vyřešen načtením pouze nejnovějšího datového souboru (historie by se pak určovala pomocí údaje `DatumZahajeniCinnosti`). Alternativně by mohl být problém vyřešen redukcí historických dat na nezbytně nutné položky nebo ukládáním pouze souhrnných hodnot pro každý kraj v daném čtvrtletí.

### 1.2 Příprava dat

Ze zkoumání zadaných dotazů a struktury dat vyplynulo, že některé atributy dat nebudou použity a tedy je možné je vypustit. Pro první dotaz **skupiny A** jsou v záznamech potřebné pouze údaje o kraji a okresu, oboru péče. Pro vyhodnocení dotazu jsou potřebné pouze aktuální data.

Druhý dotaz **skupiny A** využívá pro práci z daty pouze položku obor péče. Jako jediný zpracovává ale i historické údaje, proto je pro něj potřeba načíst i starší datové soubory o poskytovatelích zdravotních služeb. Protože využívá pouze položku obor péče, bylo by možné historické záznamy ukládat pouze jako trojice ("`ZdravotnickeZarizeniId`", "`OborPece`", "`datum vytvoření datového souboru, do kterého záznam patří`") nebo rovnou jako celkové počty poskytovatelů oborů pro každé čtvrtletí.

### 1.3 Uložení dat do databáze

Skript `download.py` i k vytvoření Mongo klienta pro připojení ke cloudové databázi a následné vytvoření databáze. Připravená data z obou zadaných zdrojů nahrává do databáze, pokud není použit argument `--local`.

## 2 Implementovaný systém pro získání, ukládání a zpracování dat