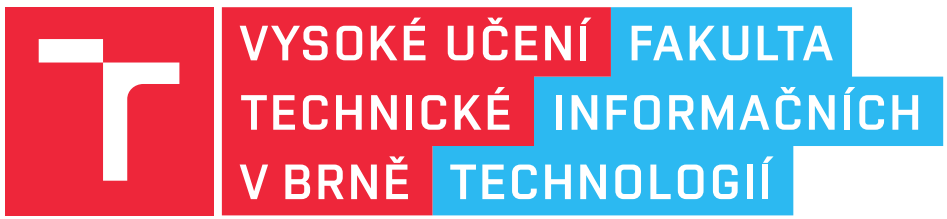
Integrace, vizualizace a dolování z dat zemí světa

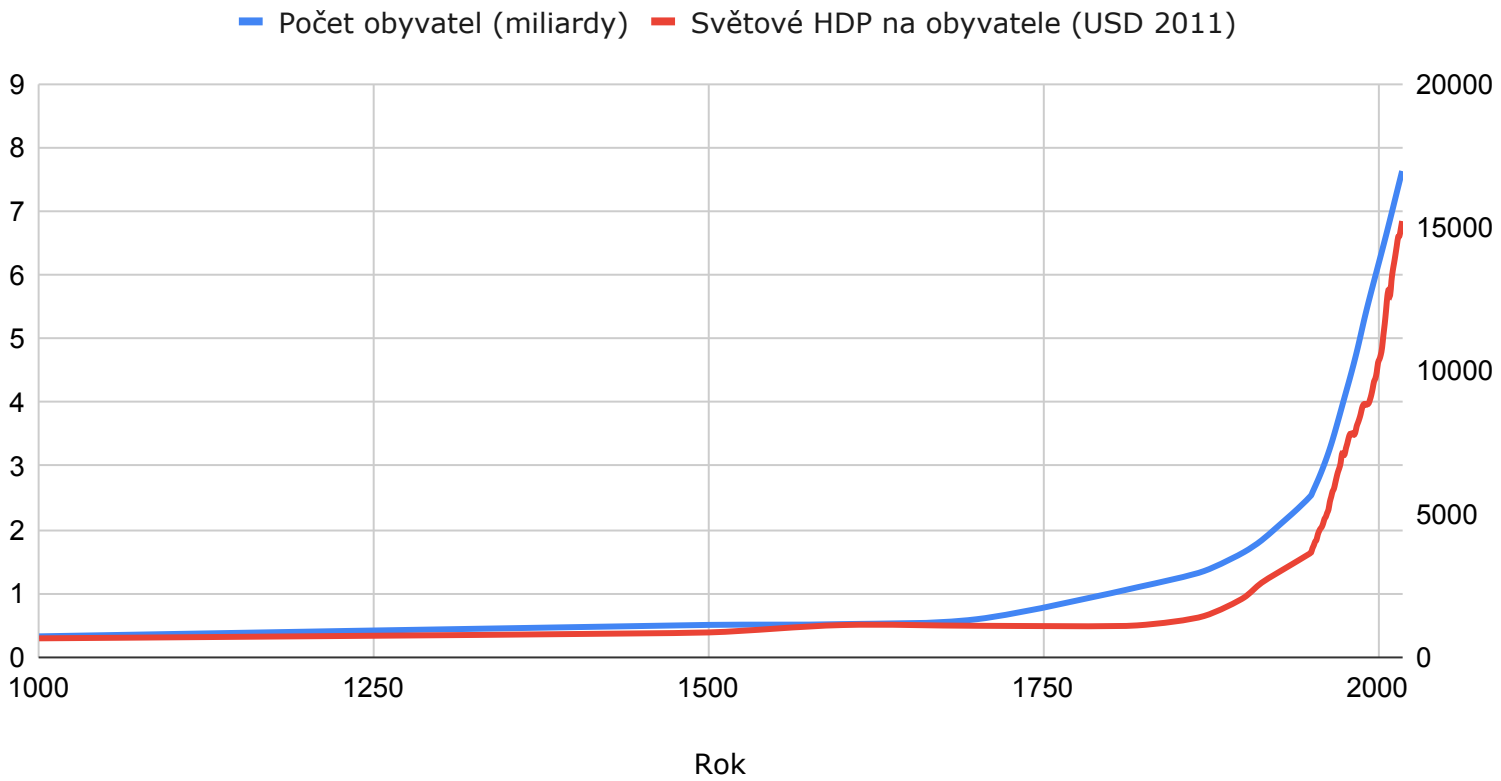
Bc. Vladimír Dušek

Vedoucí: Ing. Vladimír Bartík, Ph.D.



20. června 2022

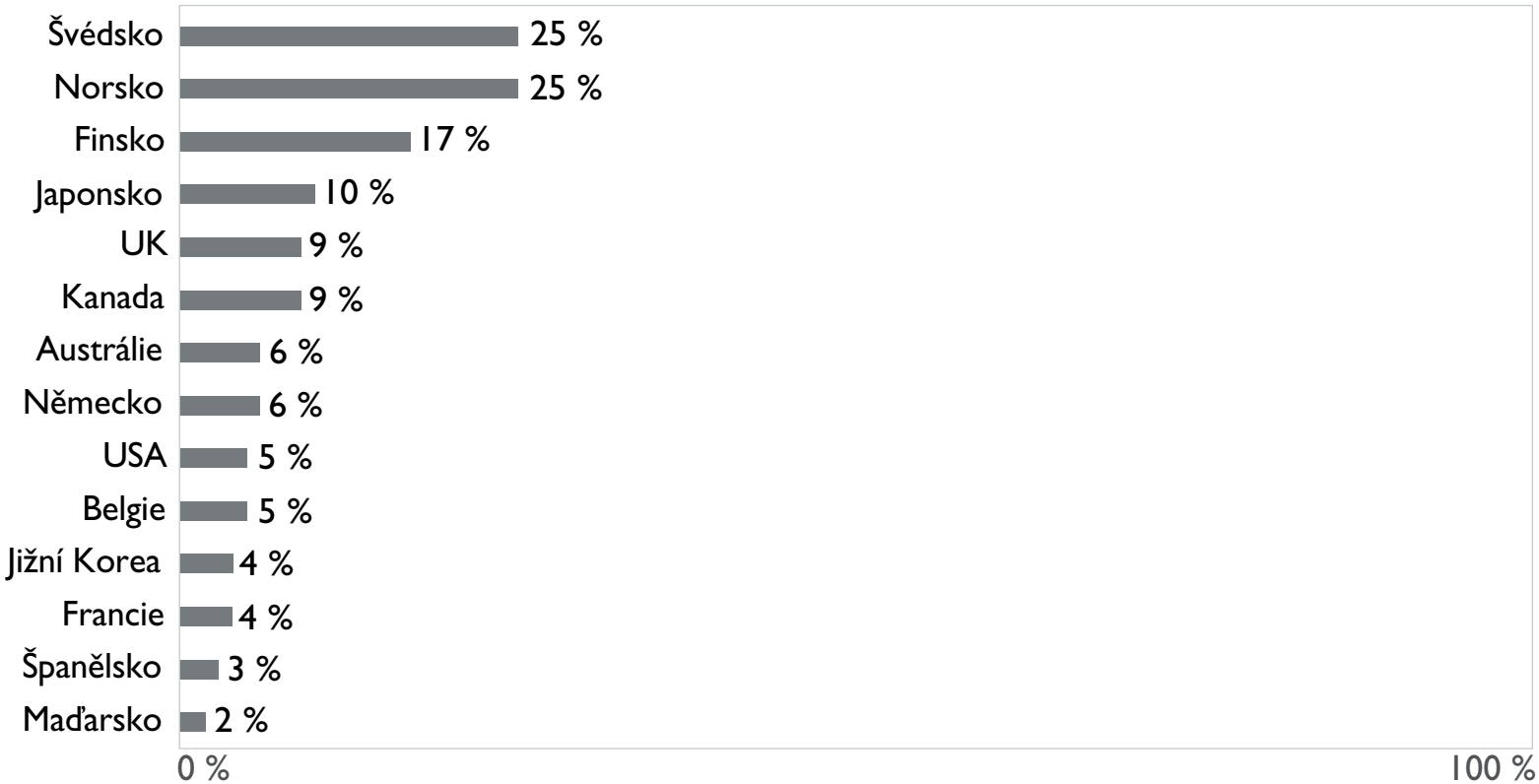
- Zvýšit povědomí o tom, jak roste životní úroveň.



Obrázek: Vývoj světové populace a světového HDP na obyvatele od roku 1000 do současnosti.[1]

[1] https://rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2020

- Většina lidí má velmi mylné představy.



Obrázek: Poměr osob odpovídajících správně na otázku: „Jak se během posledních 20 let změnil podíl světové populace žijící v chudobě?" Na výběr bylo ze tří možností: zvýšil se na dvojnásobek, zůstal stejný, **snížil se na polovinu**.[2]

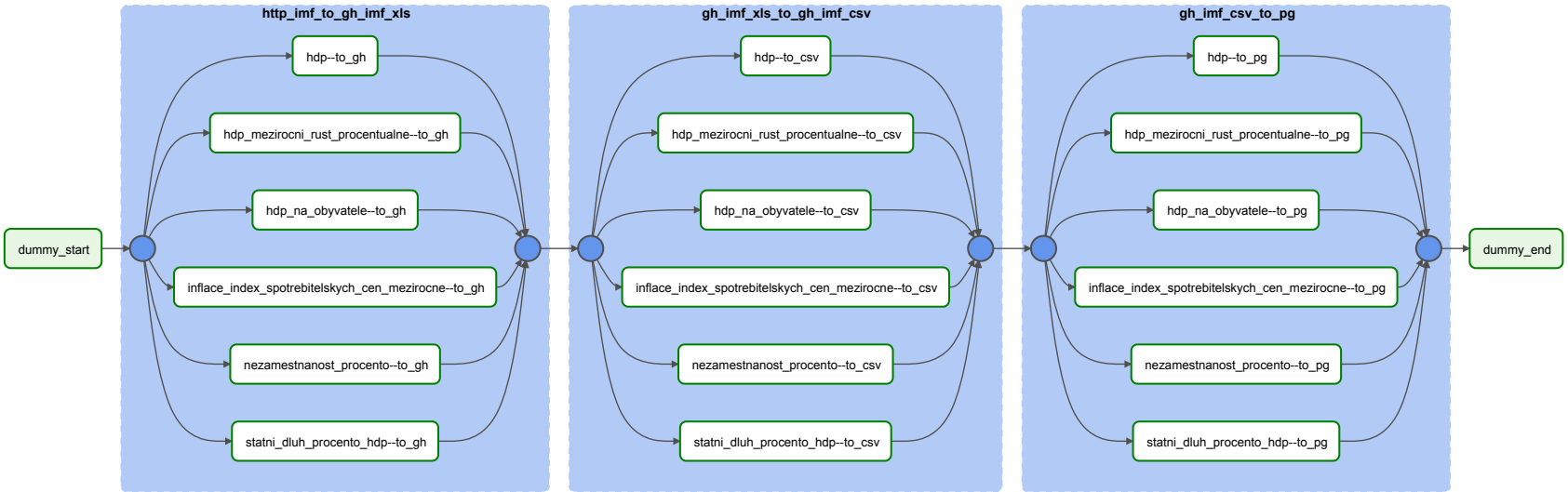
[2] Z knihy Faktomluva od Hanse Roslinga.

- Vytvořit webovou aplikaci pro prezentaci dat o globálních trendech ve světě.

Dílčí cíle

1. Prozkoumat otevřené zdroje pro získání dat
2. Databáze
3. ETL procesy pro stahování a zpracování dat
4. Webová aplikace pro prezentaci dat
5. Dolování ze získaných dat

- Otevřené datové zdroje: Světová banka, Mezinárodní měnový fond, Fraserův institut, Eurostat, OECD atd.
- *Workflows* pro Apache Airflow
- Celkem stahováno 55 indikátorů ze 3 zdrojů

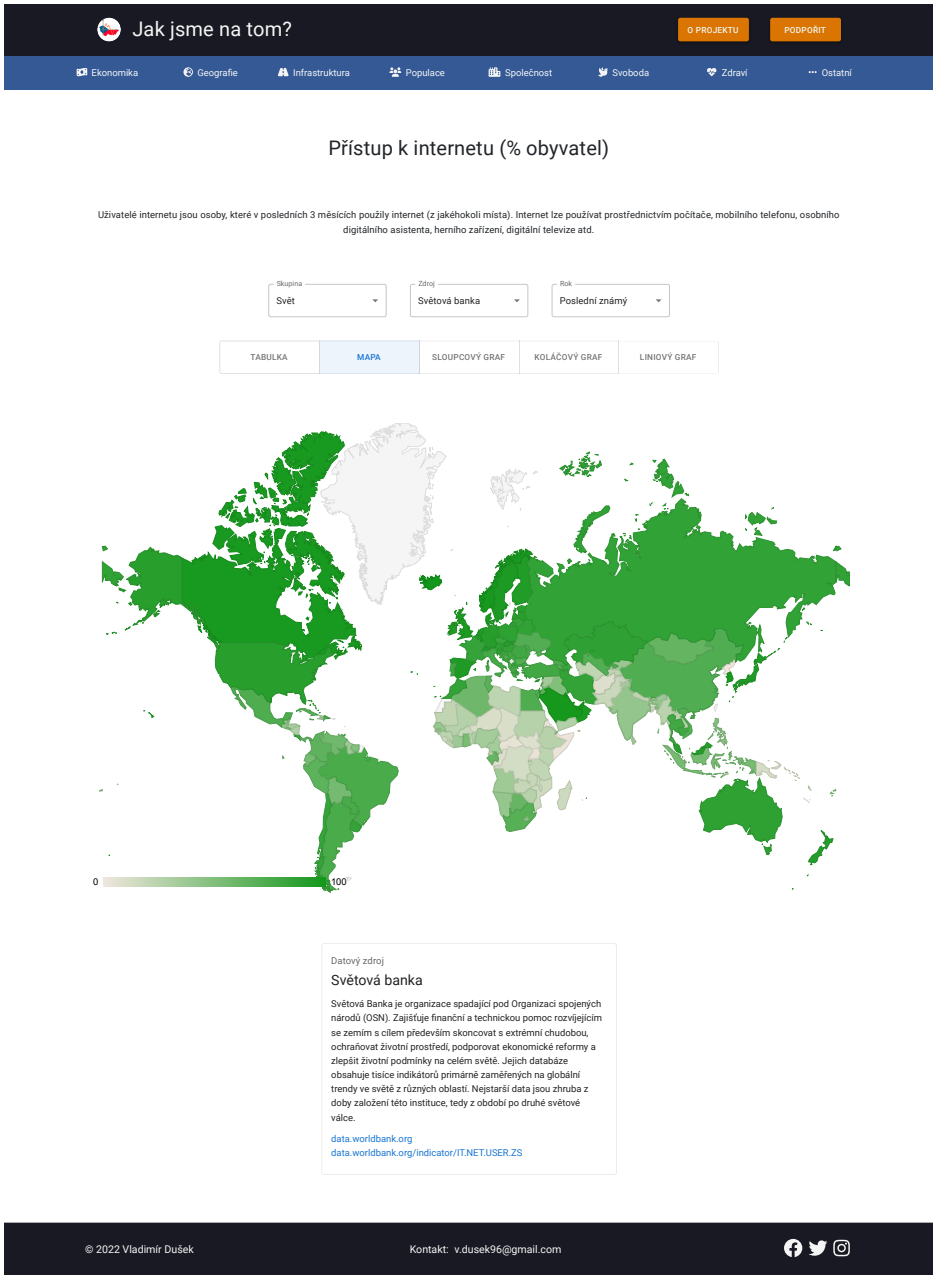


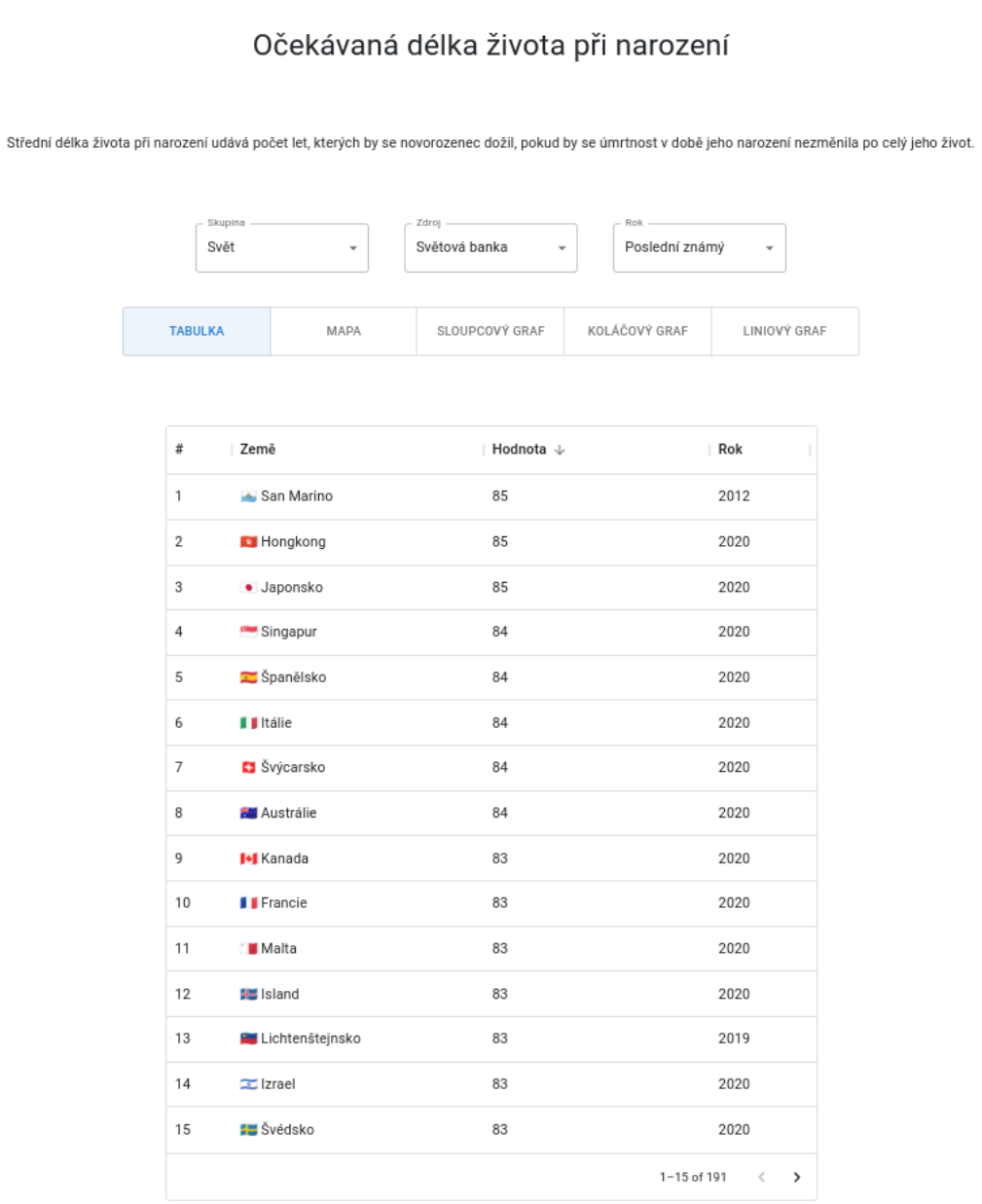
Obrázek: Ukázkové *workflow* pro stahování dat z Mezinárodního měnového fondu.

Implementace aplikace

- Databáze
  - PostgreSQL
- Serverová aplikace (API)
  - Python, framework FastAPI
- Webová aplikace
  - ReactJS
  - Uživatelské prvky – Material-UI
  - Vizualizace – Google Charts
- Aplikace je nasazena v Google Cloudu a dostupná na doméně `jakjsmenatom.cz`.

Ukázka webové aplikace 1/4





Očekávaná délka života při narození

Střední délka života při narození udává počet let, kterých by se novorozenec dožil, pokud by se úmrtnost v době jeho narození nezměnila po celý jeho život.

Skupina: Svět | Zdroj: Světová banka | Rok: Poslední známý

TABULKA | MAPA | SLOUPCOVÝ GRAF | KOLÁČOVÝ GRAF | LINIOVÝ GRAF

#	Země	Hodnota	Rok
1	San Marino	85	2012
2	Hongkong	85	2020
3	Japonsko	85	2020
4	Singapur	84	2020
5	Španělsko	84	2020
6	Itálie	84	2020
7	Švýcarsko	84	2020
8	Austrálie	84	2020
9	Kanada	83	2020
10	Francie	83	2020
11	Malta	83	2020
12	Island	83	2020
13	Lichtenštejnsko	83	2019
14	Izrael	83	2020
15	Švédsko	83	2020

1–15 of 191

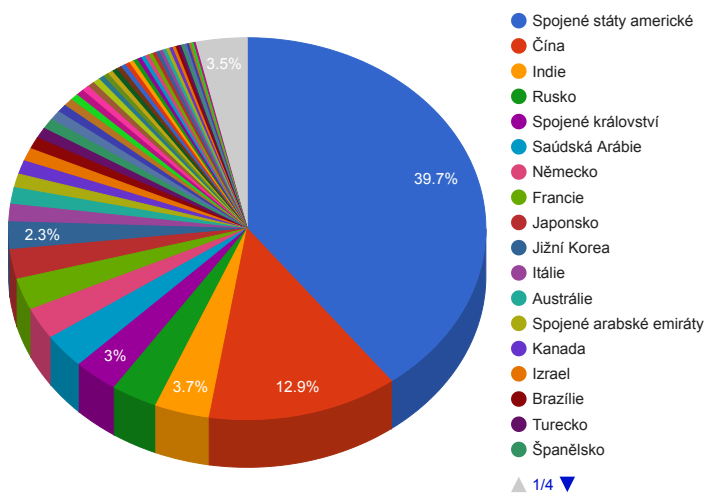
Armádní výdaje

Údaje o vojenských výdajích SIPRI vycházejí z definice NATO, která zahrnuje veškeré běžné a kapitálové výdaje na ozbrojené síly, včetně mírových sil, ministerstva obrany a další vládní agentury zapojené do obranných projektů, polovojenské síly, pokud jsou považovány za vycvičené a vybavené pro vojenské operace, a vojenské kosmické aktivity. Tyto výdaje zahrnují vojenský a civilní personál, včetně výsluhových důchodů vojenského personálu a sociálních služeb pro personál; provoz a údržbu; nákupy; vojenský výzkum a vývoj; a vojenskou pomoc (ve vojenských výdajích dárcovské země). Nezahrnují se výdaje na civilní obranu a běžné výdaje na dřívější vojenské aktivity, např. na dávky veteránům, demobilizaci, konverzi a likvidaci zbraní. Tuto definici však nelze použít pro všechny země, protože by to vyžadovalo mnohem podrobnější informace, než jaké jsou k dispozici o tom, co je zahrnuto ve vojenských rozpočtech a mimorozpočtových položkách vojenských výdajů. (Například vojenské rozpočty mohou, ale nemusí zahrnovat civilní obranu, zálohy a pomocné síly, policejní a polovojenské síly, síly dvojího určení, jako je vojenská a civilní policie, vojenské naturální dotace, důchody pro vojenský personál a příspěvky na sociální zabezpečení placené jednou částí vlády druhé).

Skupina: Svět | Zdroj: Světová banka | Rok: Poslední známý

TABULKA | MAPA | SLOUPCOVÝ GRAF | KOLÁČOVÝ GRAF | LINIOVÝ GRAF

Jednotka: $

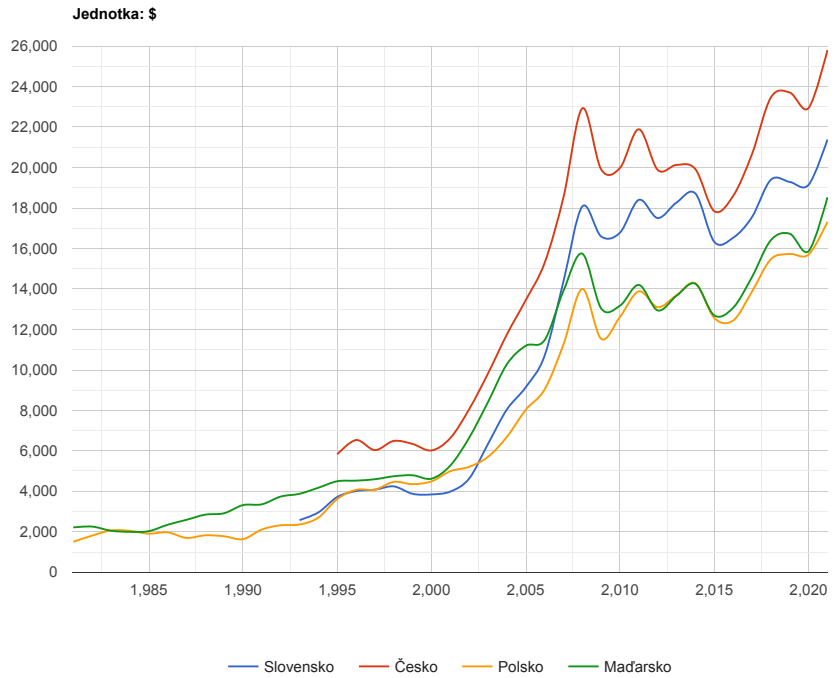


HDP (hrubý domácí produkt) na obyvatele

HDP na obyvatele je hrubý domácí produkt dělený počtem obyvatel v polovině roku. HDP je součet hrubé přidané hodnoty všech rezidentských výrobců v ekonomice zvýšený o případné daně z produktů a snížený o případné dotace, které nejsou zahrnuty v hodnotě produktů. Počítá se bez odečtení odpisů vyrobených aktiv nebo vyčerpání a znehodnocení přírodních zdrojů. Údaje jsou uvedeny v běžných amerických dolarech.

Skupina: Visegrádska čtyřka | Zdroj: Mezinárodní měnov... | Rok od: 1981 | Rok do: Poslední známý

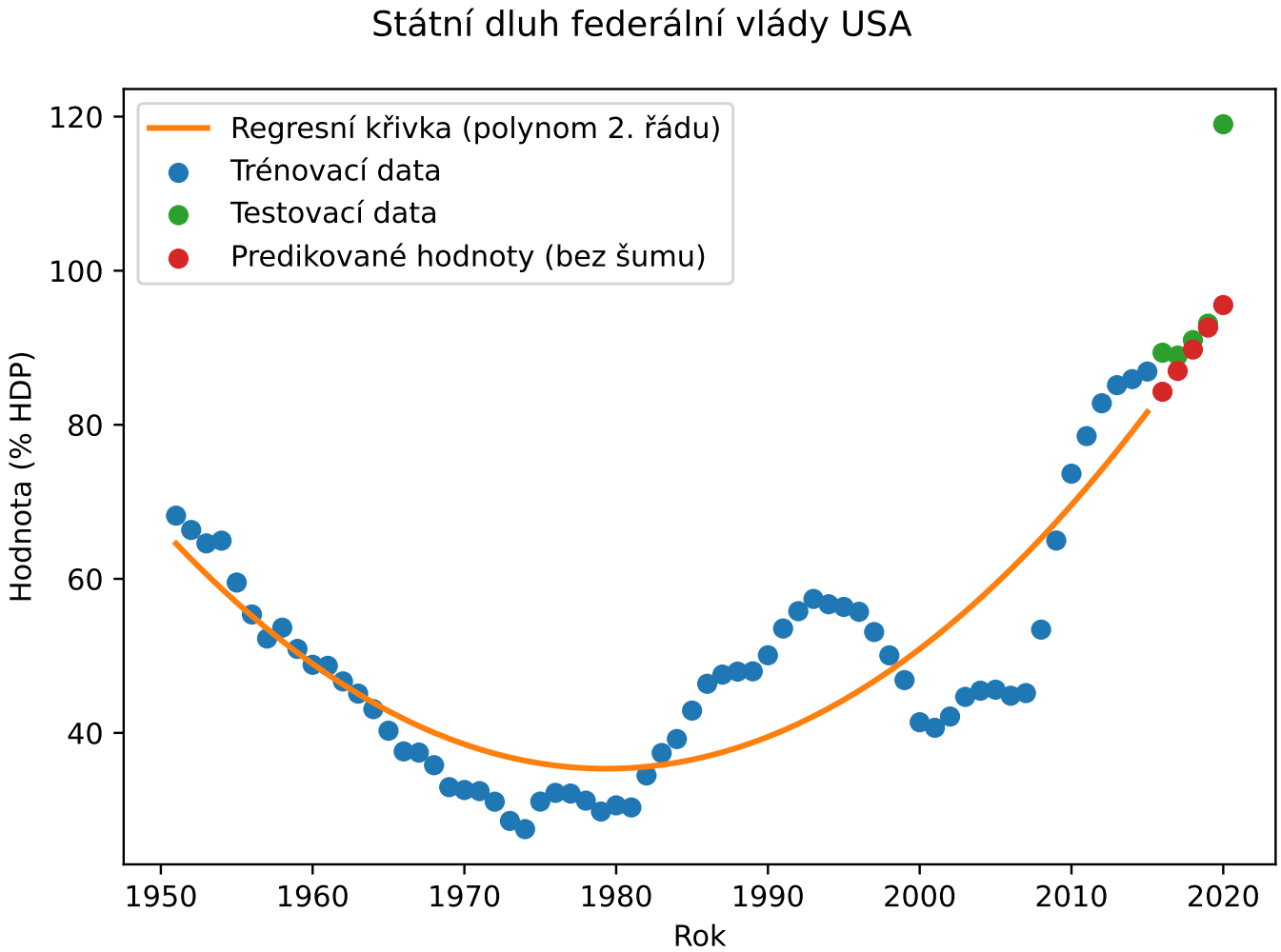
TABULKA | MAPA | SLOUPCOVÝ GRAF | KOLÁČOVÝ GRAF | LINIOVÝ GRAF



Dolování z dat

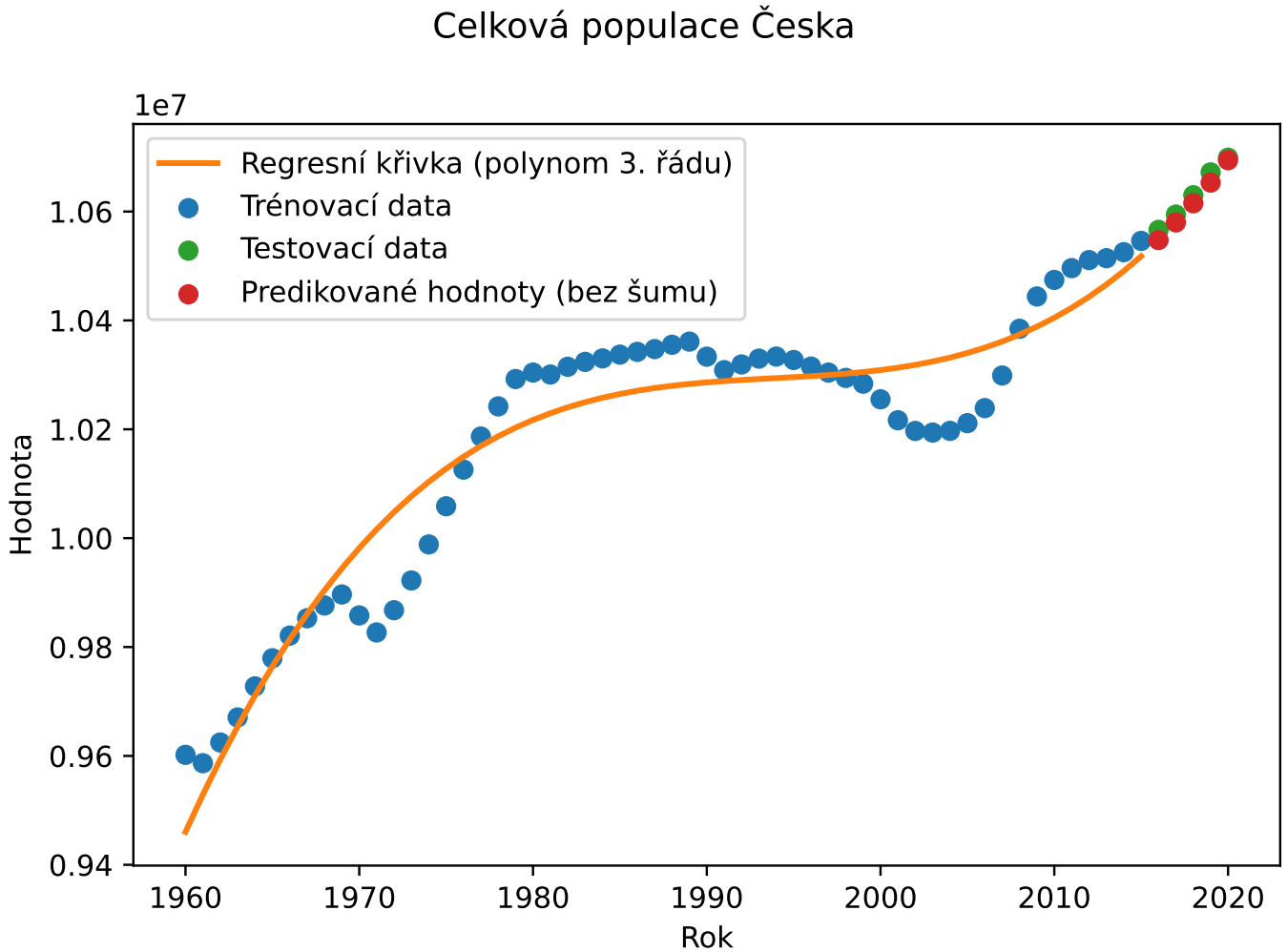
1. Regrese
   - Predikce vývoje indikátorů do budoucnosti.
2. Shlukování
   - Seskupování zemí na základě různých faktorů (indikátorů).

Ukázka dolování – regrese 1/2



Obrázek: Predikce indikátoru „státní dluh federální vlády USA (% HDP)".

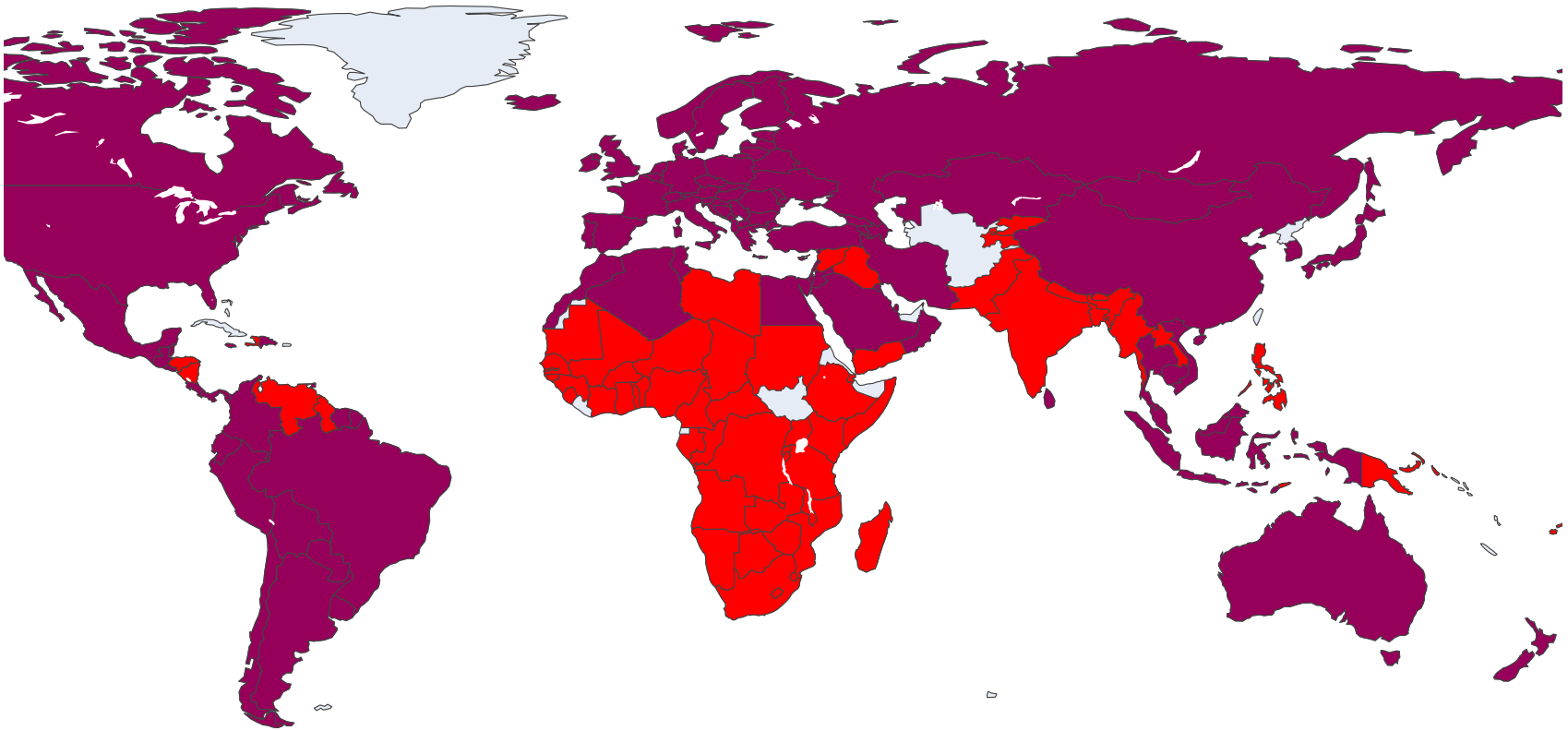
Ukázka dolování – regrese 2/2



Obrázek: Predikce indikátoru „celková populace Česka".

Ukázka dolování – shlukování 1/2

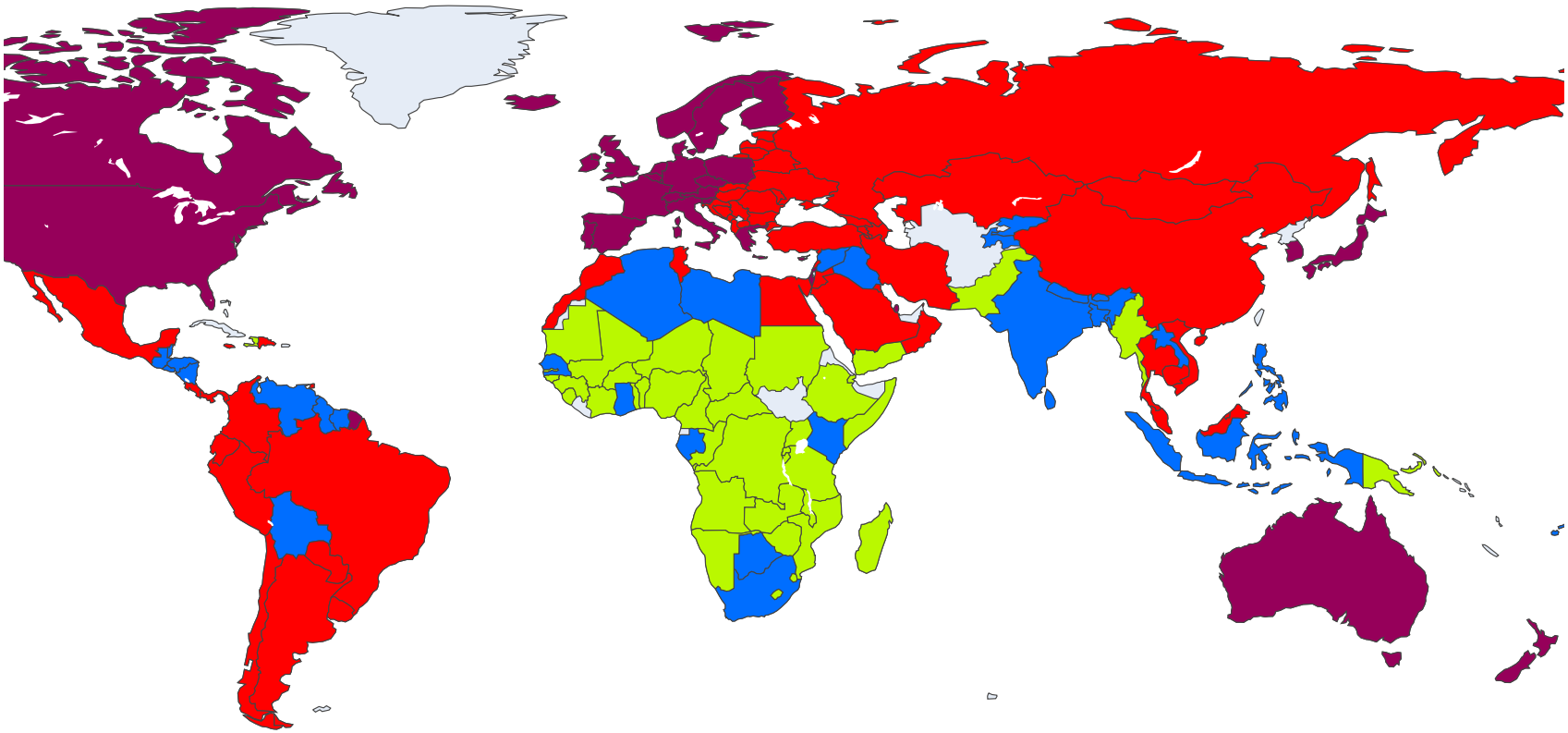
- Shlukování na základě 10 indikátorů z oblasti kvality života (HDP na obyvatele, očekávaná délka života, cestovní ruch, přístup k elektřině a internetu, inflace, . . .).



Obrázek: Shlukování pomocí K-medoids do 2 shluků.

Ukázka dolování – shlukování 2/2

- Shlukování na základě 10 indikátorů z oblasti kvality života (HDP na obyvatele, očekávaná délka života, cestovní ruch, přístup k elektřině a internetu, inflace, . . .).



Obrázek: Shlukování pomocí K-medoids do 5 shluků.

Shrnutí

- Webová aplikace „Jak jsme na tom?“.[3]
- Datové integrace v Apache Airflow.
- Dolování ze získaných dat.

[3] `https://jakjsmenatom.cz`

Shrnutí

- Webová aplikace „Jak jsme na tom?“.[4]
- Datové integrace v Apache Airflow.
- Dolování ze získaných dat.

Další kroky

- Soutěž „Společně otevíráme data 2022“ nadace OSF[5], která oceňuje nejlepší neziskové aplikace postavené nad otevřenými daty v Česku.
- Pokračovat ve vývoji a udržovat aplikaci.

[4] https://jakjsmenatom.cz

[5] https://osf.cz/programy/ziva-demokracie/nas-stat-nase-data/soutez-spolecne-otevirame-data-2021

Děkuji za pozornost!

Otázka od oponenta 1/3

1. Jak často probíhá získávání dat, plnění (aktualizace) databáze a následné dolování z dat?

Otázka od oponenta 1/3

1. Jak často probíhá získávání dat, plnění (aktualizace) databáze a následné dolování z dat?

- Apache Airflow není nikde nasazené.
  - Datové integrace nejsou nijak plánované a jsou spouštěny ad hoc.
  - Pokud by nasazené bylo, dávalo by smysl naplánovat běh např. 2× do měsíce (vzhledem k četnosti aktualizací zdrojových dat).

Otázka od oponenta 1/3

1. Jak často probíhá získávání dat, plnění (aktualizace) databáze a následné dolování z dat?

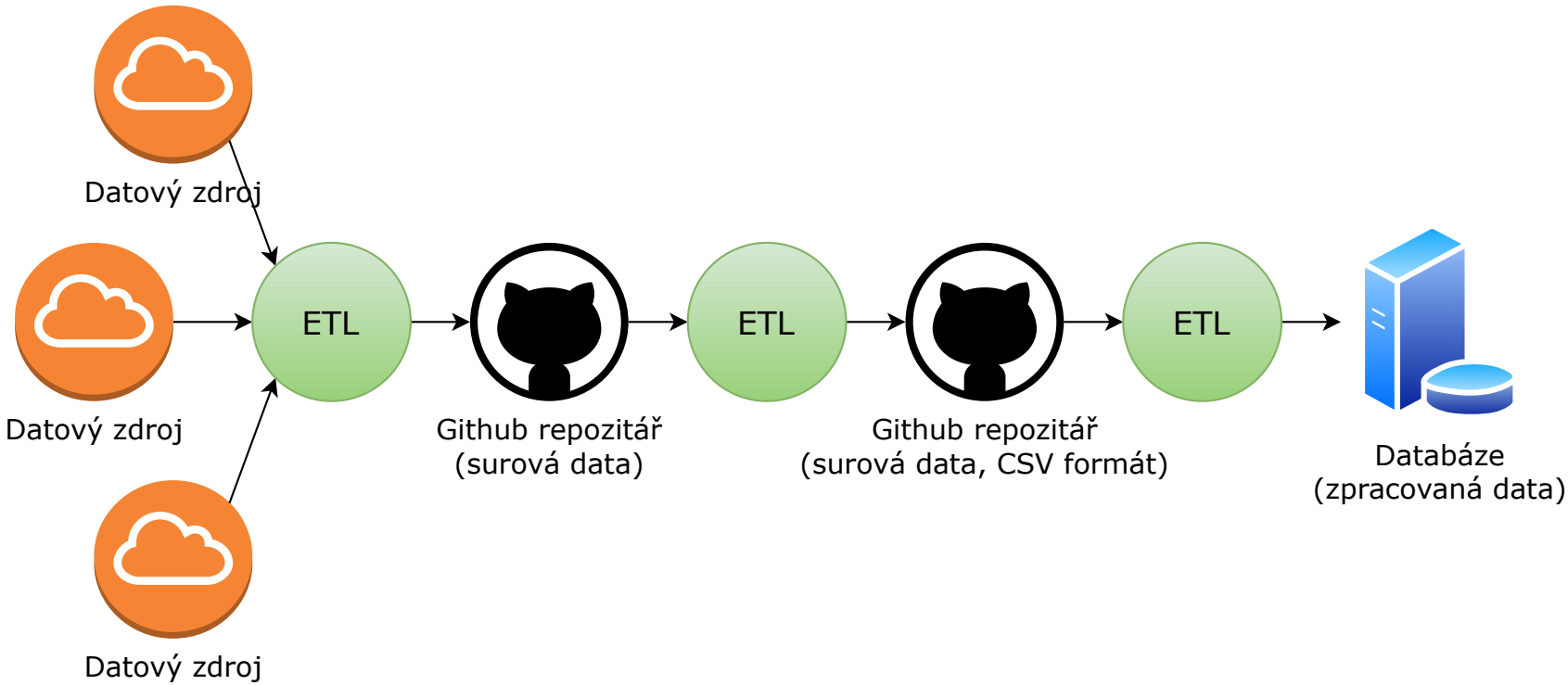
- Apache Airflow není nikde nasazené.
  - Datové integrace nejsou nijak plánované a jsou spouštěny ad hoc.
  - Pokud by nasazené bylo, dávalo by smysl naplánovat běh např. 2× do měsíce (vzhledem k četnosti aktualizací zdrojových dat).
- Dolovací skripty nejsou implementovány jako *workflows* pro Apache Airflow.
  - Nejsou nijak plánované.

Otázka od oponenta 2/3

2. **Jak dlouho tyto aktivity celkově trvají** a jak by bylo možné je optimalizovat?

2. **Jak dlouho tyto aktivity celkově trvají** a jak by bylo možné je optimalizovat?

- Každý indikátor se zpracovává ve 3 úlohách.



Obrázek: Schéma stahování, zpracování a ukládání dat.

Otázka od oponenta 2/3

2. **Jak dlouho tyto aktivity celkově trvají** a jak by bylo možné je optimalizovat?

- Je integrováno 55 indikátorů, tj. celkem 165 úloh.
- Kvůli nahrávání dat do Github repozitáře se úlohy v první a druhé úrovni musí provádět sekvenčně.
  - Pro zjednodušení uvažujme, že i úlohy třetí úrovně jsou prováděny sekvenčně.
- Jedna úloha trvá v průměru asi 20 sekund, tj. celkem 55 minut.

Otázka od oponenta 3/3

2. Jak dlouho tyto aktivity celkově trvají a **jak by bylo možné je optimalizovat**?

Otázka od oponenta 3/3

2. Jak dlouho tyto aktivity celkově trvají a **jak by bylo možné je optimalizovat**?

- *Bottleneck* řešení spočívá v Githubu, alternativně by šla použít libovolná *object storage* (např. Google Cloud Storage).
- To by umožnilo úlohy vykonávat paralelně.
- Stávající konfigurace Airflow (Celery Executor, 1 dělník) umožňuje provádět 16 úloh současně.
- Tím by bylo možné zrychlit běh datových integrací 16× (z 55 minut na asi 3,5 minuty).