

# VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

## Fakulta Informačních technologií

Projekt z UPA – Příprava a ukládání dat  
Rychlost internetového připojení

# Návrh

### Stručná charakteristika zvolené datové sady:

* Zdrojem dat bude soubor s měsíčními výsledky databáze NetMetr (konkrétně září 2020)
* Data budou stažena z odkazu <https://www.netmetr.cz/cs/open-data.html>
* Vstupní soubor je ve formátu csv. Na prvním řádku se nachází hlavička a díky ní ze souboru vyfiltrujeme užitečná data, například download\_kbit a upload\_kbit.

### Předzpracování dat pro NoSQL databázi:

* Odstranění závadných dat – s parametrem implausible nebo s tagy chyb.
* Odstranění neúplných dat, které neobsahují rychlosti stahování, odesílání nebo odezvy.

### Převod z NoSQL do relační databáze:

* Rozdělení dat z kolekce do dvou tabulek pro zamezení redundance dat – uživatelé a provedené testy rychlosti připojení.
* Tabulka s uživateli obsahuje jejich ID, prefix jejich IP adresy, technologii/typ připojení, název mobilní sítě
* Tabulka s provedenými testy obsahuje ID testu, ID uživatele, data o rychlosti připojení (stahování, odesílání, odezva), platformu a sílu signálu (pro mobilní sítě)

### Zvolené dotazy a formulace vlastního dotazu:

* Skupina A: vytvořte popisné charakteristiky pro měřené hodnoty rychlosti stahování, odesílání a odezvy pro zvoleného uživatele či sít (IP prefix); využijte krabicové grafy, histogramy atd.
  + Dotaz vyřešíme spojením tabulek uživatelů a testů a následnou filtrací dle zadaného prefixu IP adresy. Výsledné charakteristiky měření zobrazíme v krabicovém grafu.
* Skupina B: seskupte uživatele podle rychlosti jejich připojení (v intervalech) a určete společné a rozdílené vlastnosti jednotlivých skupin
  + Dotaz by byl řešen spojením tabulek uživatelů a testů, seskupením uživatelů dle nějaké charakteristiky rychlosti připojení – např. maximální nebo střední hodnota rychlosti stahování, odesílání, odezvy, nebo celkového průtoku dat při testu.
  + Sledované vlastnosti skupin mohou být síla signálu, technologie/typ připojení, název mobilní sítě nebo platforma použitého mobilního zařízení.
  + Celkový průtok dat může být počítán jako agregace rychlosti stahování a odesílání.
* Vlastní dotaz: vytvořte popisné charakteristiky pro měřené hodnoty rychlosti stahování, odesílání a odezvy na základě použité technologie (3G, 4G/LTE, LAN)
  + Dotaz vyřešíme spojením tabulek uživatelů a testů a následnou agregací sledovaných údajů do skupin dle použité technologie. Výsledné charakteristiky měření zobrazíme v krabicovém grafu.

# Technologie

### Zvolený způsob uložení dat:

* Pro uložení surových dat jsme zvolili NoSQL databázi MongoDB, protože je uživatelsky přívětivá a podporuje všechny potřebné funkce pro řešení projektu (např. agregace).
* MongoDB je dokumentová databáze, která ukládá data ve stylu formátu JSON v takzvaných kolekcích.
* Pro cílové uložení dat jsme zvolili relační databázi MySQL, se kterou již máme zkušenosti a tedy víme, jak s ní pracovat.

### Zvolený jazyk pro implementaci systému:

* Pro implementaci jsme zvolili jazyk Python 3, jelikož má kvalitní knihovny, umí dobře vykreslovat grafy a je efektivní z hlediska délky zdrojových kódů.

# Implementace a spuštění

Implementaci jsme provedli pomocí 4 skriptů v jazyce Python 3 – clear\_db.py, import\_data.py, clean\_data.py, a results.py.

Pro spuštění je tedy vyžadován jazyk Python 3 a instalace balíčků ze souboru requirement.txt, dále je zapotřebí běžící databáze MySQL a MongoDB, pro které je nutné vyplnit přihlašovací údaje k databázím v souboru settings.py.

### Clear\_db.py:

* Vymaže databázi v MongoDB.
* Příklad spuštění: clear\_db.py

### Import\_data.py:

* Uloží data z csv souboru do databáze MongoDB.
* Příklad spuštění: import\_data.py data.csv

### Clean\_data:

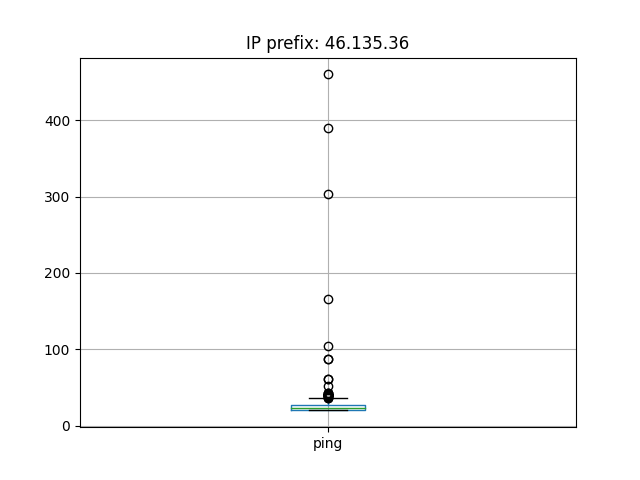
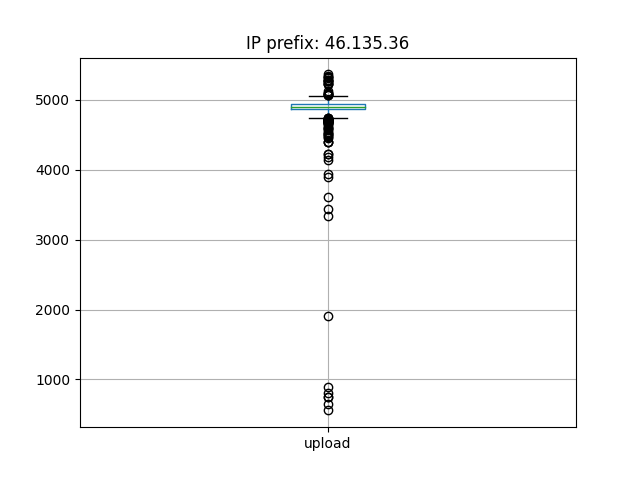
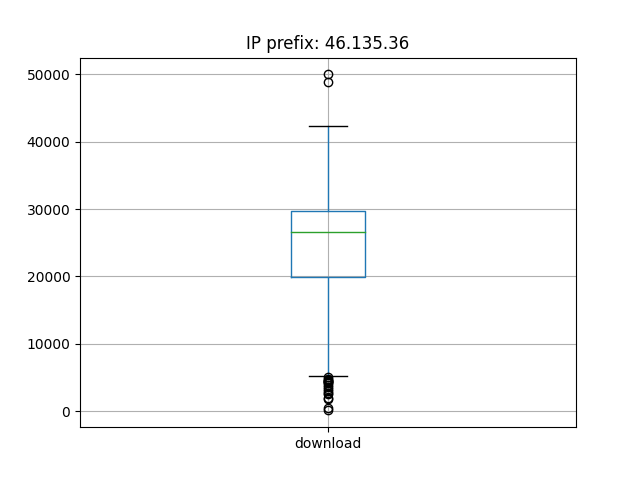
* Připojí se k MySQL databázi a převede do ní vyfiltrovaná potřebná data z MongoDB databáze.
* Data jsou do MySQL databáze nahrávaná po dávkách o velikosti 10000 položek.
* Příklad spuštění: clean\_data.py

### Results.py:

* Dle zadaných parametrů zodpoví dotaz, provede analýzu dat a zobrazí příslušné grafy.
* Příklady spuštění:
  + clean\_data.py 46.135.36
    - Zodpoví dotaz A a zobrazí grafy charakteristik rychlosti připojení pro zadaný prefix IP adresy.
  + clean\_data.py
    - Zodpoví náš vlastní dotaz a zobrazí příslušné grafy rychlosti připojení.

# Výsledky

## Dotaz A



## Vlastní dotaz

