Report - zimný semester
Ročníkový projekt
Práca s databázou s údajmi z ping a traceroute

Filip Hošták

Náplňou mojej práce počas zimného semestra bolo najmä zoznamovanie sa so samotnou databázou, ako aj základné spracovanie údajov a niekoľko vizualizácií. Bolo to moje prvé stretnutie so vzdialenou *PostgreSQL* databázou, teda bolo potrebné sa najprv oboznámiť so základmi *PostgreSQL*. Na ďalšiu prácu s údajmi som použil programovací jazyk *Python* spolu s 3 knižnicami - *psycopg2* pre prácu s *PostgreSQL* databázami, *matplotlib* na vizualizácie, *maxminddb* na prácu s databázou *IP To City Lite*.

Stručný popis databázy

Databáza sa skladá zo 7 relácií, nás najviac zaujímajú 3:

1. **hosts** - Záznam všetkých IP adries, z ktorých boli dáta čerpané, ako aj ich dátum pridania do databázy. [281 330 záznamov]

2. **ping** - Záznam všetkých meraní ping na IP adresách z hosts, relevatné údaje sú najmä IP adresa, čas merania, packet loss (v percentách) a roundtrip average, minimum a maximum. Spoľahivý záznam budeme považovať taký, ktorý má packet loss medzi 0 (vrátane) a 100 (vynímajúc) a zároveň má čas merania najskôr v roku 2009. Záznamy sa zvykli robiť každý deň / IP adresa. [563 605 275 záznamov]

3. **topology** - Záznam všetkých meraní traceroute na IP adresách z hosts, relevatné údaje sú najmä IP adresa, čas merania, roundtrip, počet hopov, exit status (jeden znak), ako aj samotná cesta v podobe zoznamu IP adries. Spoľahlivý záznam budeme v niektorých prípadoch považovať také, ktoré majú exit status iný ako 'E' (Error) a v iných prípadoch len také, ktoré majú výlučne exit status 'C' (Complete). Pre všetky spoľahlivé záznamy taktiež musí platiť, že ich čas merania je najskôr v roku 2009. Záznamy sa spočiatku robili každý deň / IP adresa, neskôr každý týždeň / IP adresa. [71 187 905 záznamov]

Práca s databázou

V databáze sa vyskytujú určité nezrovnalosti, z ktorých najzávažnejšou je tá, že viacero záznamov má čas merania spred roku 2009 - a to napriek tomu, že dáta sa začali merať a teda aj zapisovať do databázy až v roku 2009. To zároveň odôvodňuje našu predtým spomínanú podmienku pre spoľahlivý záznam. Pri množstve záznamov som taktiež musel brať ohľad na to, ako dlho sa bude daný SQL príkaz vykonávať - neraz som na danú úlohu vymyslel príkaz, ktorý nebol hotový ani po 2 hodinách. Našťastie v drvivej väčšine takýchto prípadov sa mi podarilo prísť na efektívnejší príkaz, ktorý bol hotový za menej ako 30 minút.

Čo som robil s databázou

Hľadanie aspoň 24h intervalov bez spoľahlivých dát

Vytvorenie procedúry, ktorá hľadá všetky aspoň 24 hodín dlhé intervaly bez spoľahlivých záznamov. Verzia tejto procedúry pre relácie ping a topology trvá približne 15 minút.

Spolupráca s databázou *IP to City Lite*

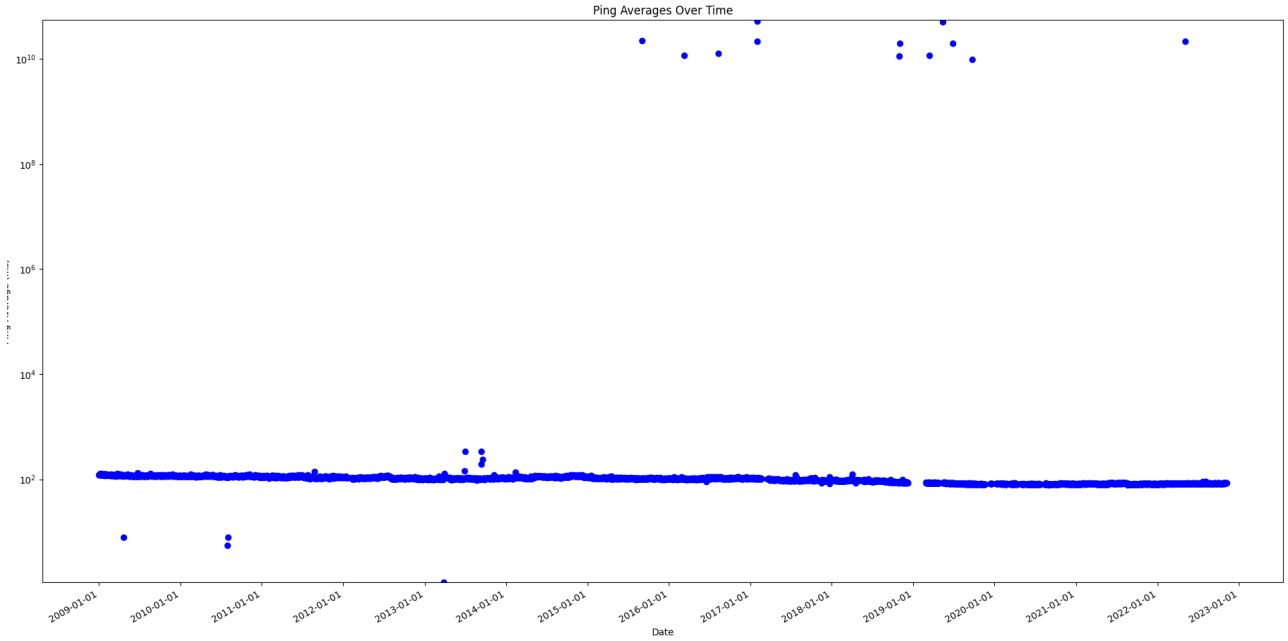
Databáza *IP to City Lite* má pre každú IP adresu záznam, ktorý obsahuje daný štát aj mesto, kde sa počítač s danou IP adresou nachádza. Bohužiaľ údaje nemusia byť aktuálne vzhľadom na to, že mnohé IP adresy sa postupom času presúvajú na iné počítače.

Vizualizácia priemerov pingu

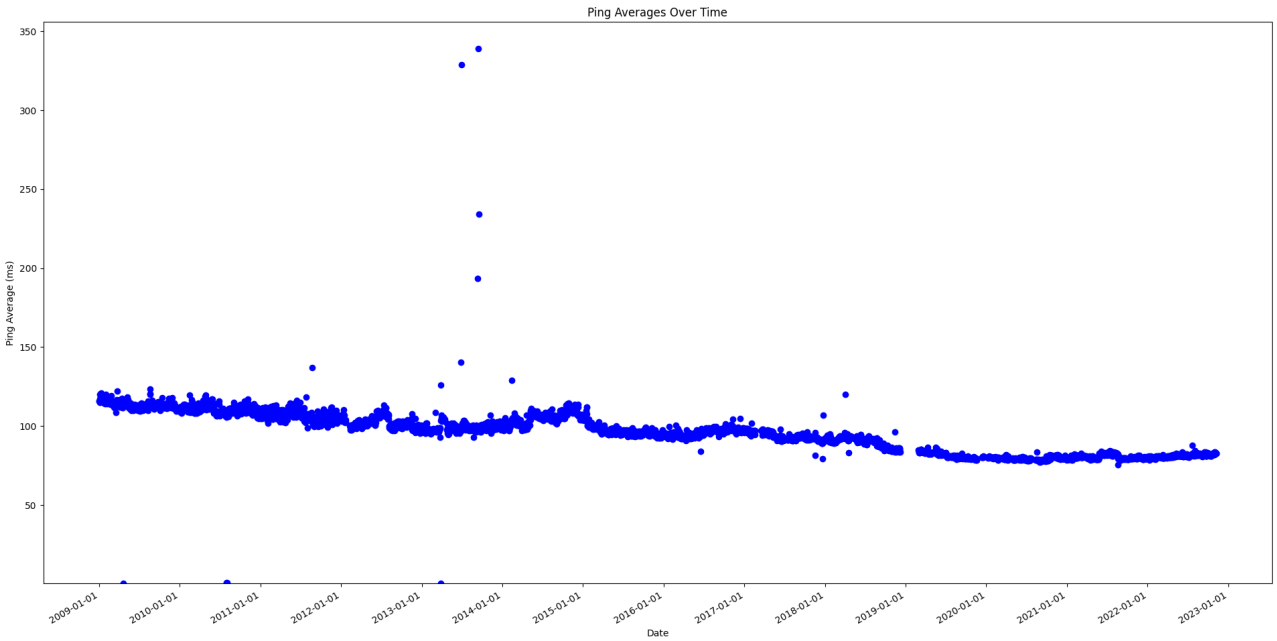
Máme viacero vizualizácií - pre každú sme kládli trochu iné podmienky, čo sa týka celkovej spoľahlivosti dát.

Všetky spoľahlivé záznamy

Všetky spoľahlivé záznamy pre všetky IP adresy, ktoré majú aspoň 2000 spoľahlivých záznamov

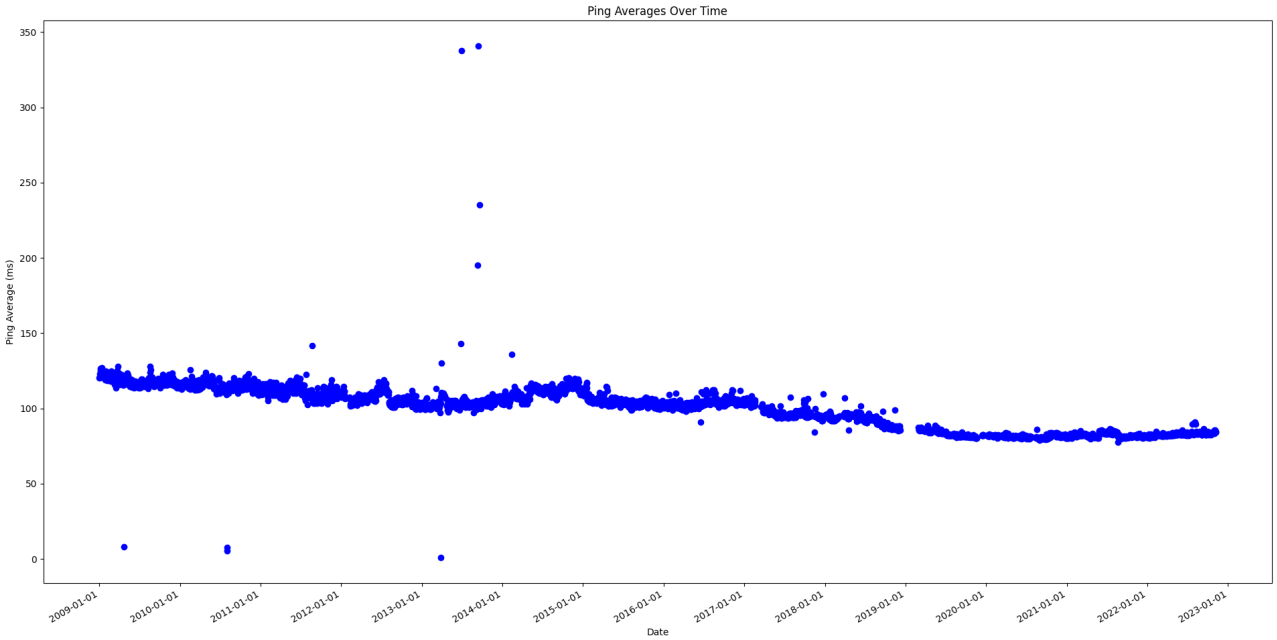


Obr. 1: Logaritmický graf priemerov ping_rttavg pre dané dni.



Obr. 2: Linerány graf priemerov ping_rttmin pre dané dni.

Všetky spoľahlivé záznamy pre dni s priemernou deviáciou meraní maximálne 19 ms



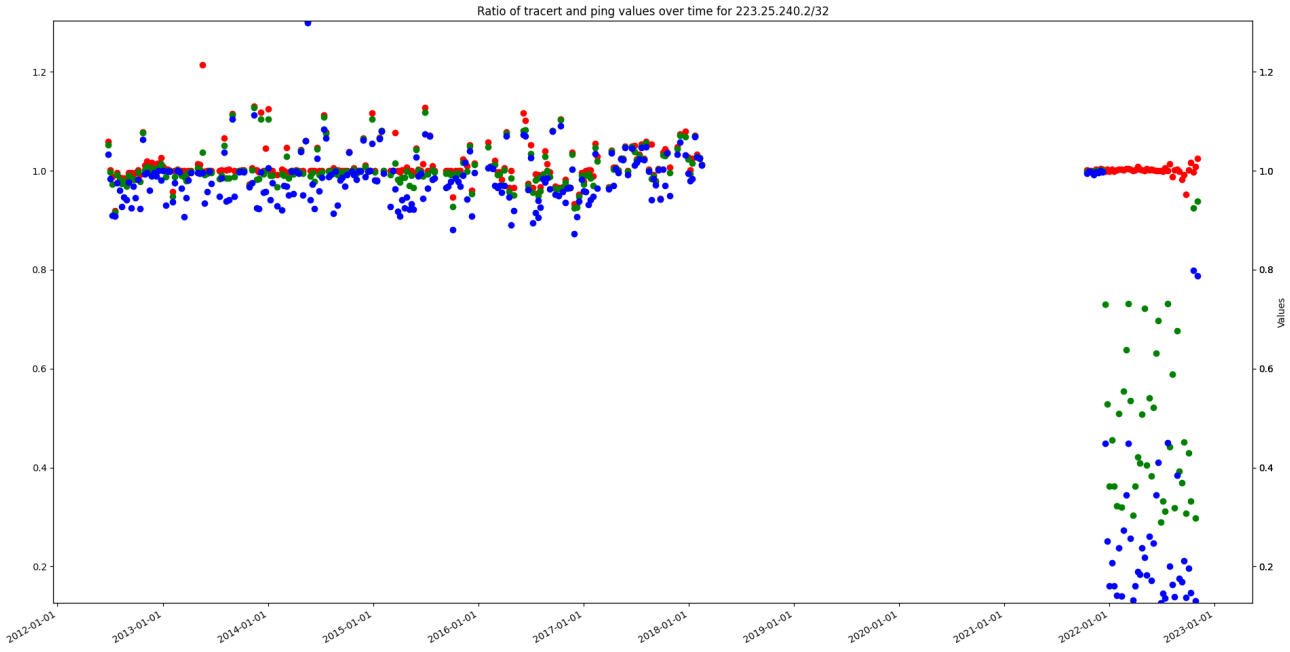
Obr. 3: Logaritmický graf priemerov ping_rttavg pre dané dni.

Súvislosť údajov v reláciách ping a topology

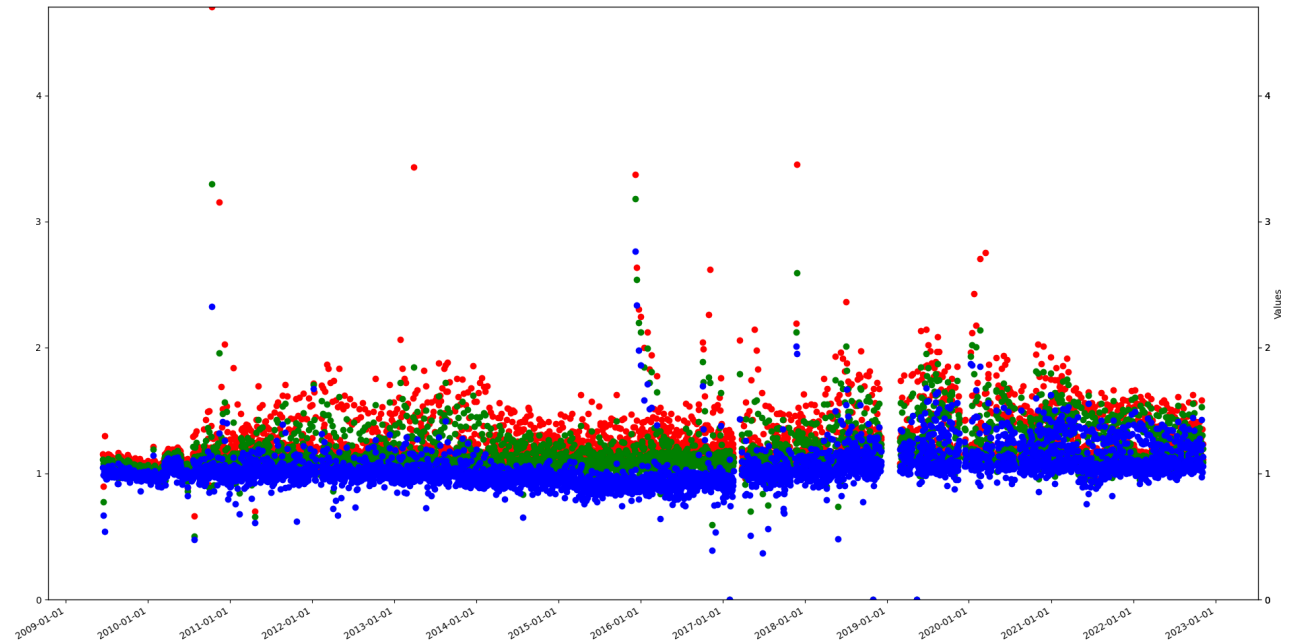
Obidve relácie obsahujú rovnakú informáciu - za akú dobu sa náš centrálny počítač dostal k cieľu. Relácia ping obsahuje údaje z programu ping, ktorý používa UDP (zvyčajne rýchlejší) a topology obsahuje údaje z programu tracert, ktorý používa TCP (zvyčajne pomalší). Pre verziu pracujúcu s údajmi 1 IP adresy sme si zobrali dni, kedy bolo úspešné meranie v ping aj tracert a urobili sme pomer priemernej dĺžky roundtripu v tracert za ten daný deň ku priemernej

hodnote rttmin, rttavg, rttmax z relácie ping za ten daný deň. Vo verzii, ktorá vizualizuje dáta pre všetky IP adresy, sme si najprv našli všetky spoľahlivé IP adresy a potom sme postupovali podobne, ako vo verzii pre 1 IP adresu s rozdielom, že priemery údajov za deň boli brané za všetky IP adresy, ktoré v daný deň mali úspešné tracert (flag 'C') aj ping meranie.

Vizualizácie (červenou pomer tracert roundtrip/ping rttmin, zelenou rttavg, modrou rttmax)



Obr. 4: Vizualizácia pre 1 IP adresu.



Obr. 5: Vizualizácia pre všetky IP adresy.