VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

http://fit.vut.cz

Projektová praxe 2021/2022

Extrahovanie znalostí o koncových zariadeniach v sieti

Adam Zvara*

Abstrakt

Táto práca je zameraná na extrakciu znalostí o koncových zariadeniach pri monitorovaní sieťových tokov. Analyzovať chovanie konkrétnej IP adresy alebo podsiete predstavuje prechádzanie veľkého množstva údajov, čo je obzvlášť na veľkých sietiach časovo a výpočetne náročné. Pre tento účel je vhodné vytvoriť agregovaný pohľad na jednotlivé koncové stanice v sieti, ktorý poskytne prehľad o tom, aké služby stanica využívala alebo poskytovala. Extrahované informácie sú následne uložené vo forme binárneho súboru s využitím dátovej kompresie, ktorého štruktúra umožňuje rýchle vyhľadávanie potrebných informácií.

Kľúčové slová: Monitorovanie siete, zber dát, tok, záznam, agregácia, extrahovanie, pohľad, binárny súbor

Priložené materiály: *N/A*

*xzvara01@stud.fit.vutbr.cz, *Fakulta informačních technologií, Vysoké učení technické v Brně*

1. Úvod

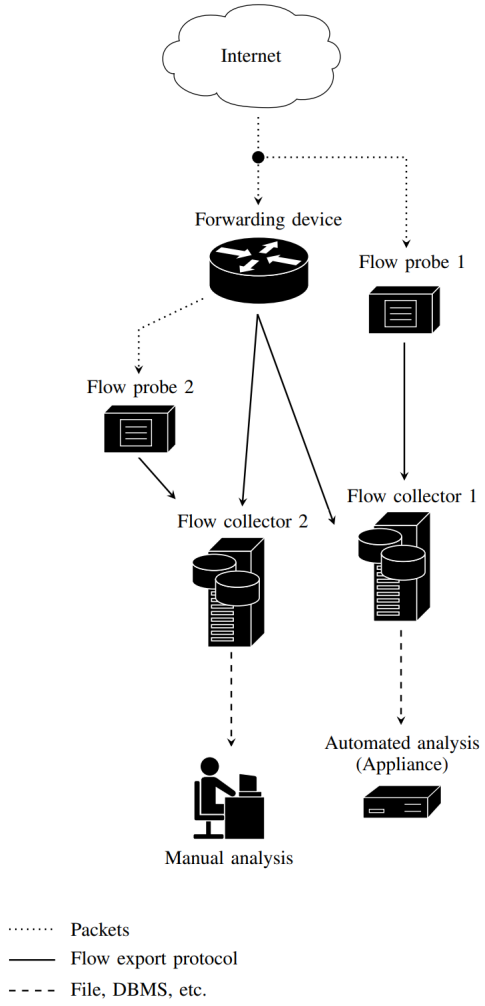
Monitorovanie siete predstavuje jednu zo základných metód analýzy siete. V súčasnej dobe spolu medzi sebou komunikuje značné množstvo zariadení. Získanie prehľadu nad prenosom dát v sieti je vhodné z pohľadu analýzy potenciálnych bezpečnostných incidentov alebo identifikácie problémov, ktoré sa na sieti môžu vyskytovať.

Ku základným protokolom na monitorovanie siete patria technológie NetFlow [3] a IP Flow Information Export (IPFIX). Ich princíp je založený na analyzovaní sieťových tokov prechádzajúcich cez monitorovanú sieť a extrahovaní štatistík komunikácií medzi užívateľmi. Za zmienku stojí fakt, že tieto technológie zvyčajne zbierajú iba štatistiky o jednotlivých komunikáciách (narozdiel od iných spôsobov, pri ktorých sa uchováva aj obsah na úrovni paketov). Získané štatistiky sú následne uložené do dátového úložiska zvaného kolektor, z ktorého môžu byť spätne manuálne analyzované alebo v nich môžu byť automaticky detekované významné sieťové udalosti. Obrázok 1 popisuje typickú architektúru zloženú zo sondy, ktorá analyzuje prenos dát na sieti a záznamy o tokoch odosiela na exportér.

Základný pojem, na ktorom sú technológie NetFlow a IPFIX založené je sieťový tok. Sieťový tok je definovaný (podľa štandardu RFC 7011 [1]) ako postupnosť paketov prechádzajúca bodom pozorovania počas určitého časového intervalu, kde všetky pakety patriace rovnakému toku majú určitú množinu spoločných vlastností. Tieto vlastnosti sú väčšinou položky získané priamo z transportných hlavičiek paketov alebo vlastnosti z nich odvodené. Existujú spoločné vlastnosti pre všetky pakety rovnakého toku (napr. zdrojová/cieľová IP adresa, zdrojový/cieľový port, číslo protokolu) a vlastnosti vypočítané, odvodené alebo prítomné len v určitých paketoch (napr. celkový počet prenesených bytov, doba kedy bol videný prvý a posledný paket).

Takýto pohľad na jednotlivé toky medzi koncovými stanicami hovorí o tom, aké služby stanica poskytuje alebo využíva a s akou intenzitou. Pomocou toho sa vieme na sieť pozrieť ako na celok a určiť jej vlastnosti ako sú napríklad najčastejšie využívané služby alebo najviac aktívne zariadenia. Na druhej strane

sa vieme zamerať na aktivitu samotných zariadení. Sme schopní určiť, o aký druh zariadenia sa jedná, kedy je aktívne, aké služby využíva alebo poskytuje, s koľkými zariadeniami komunikovalo, koľko paketov/bajtov bolo prenesených v rámci jeho komunikácie. Na základe týchto údajov si dokážeme predstaviť chovanie zariadenia a v prípade, že by sa neočakávane zmenilo, vieme určiť, v akom bode to nastalo.



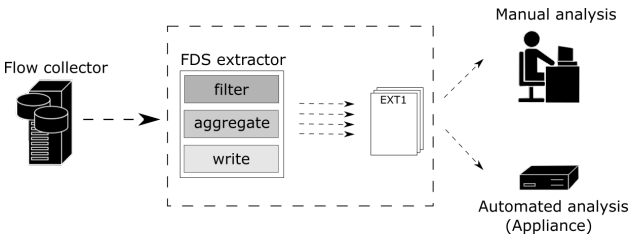
Obrázok 1. Princíp technológie NetFlow/IPFIX [4].

Ako bolo spomenuté v úvode práce, jeden z dôvodov analyzovania komunikácií v sieti je kontrola bezpečnosti. Pri vyšetrovaní bezpečnostného incidentu je vhodné najprv okrajovo lokalizovať zariadenie a čas, v ktorom došlo ku podozrivej aktivite. Na to nám slúžia informácie z kolektoru, pomocou ktorých môžeme pomerne rýchlo nájsť podozrivú stanicu a jej správanie sa v čase. Problémom je množstvo uložených dát, ktoré je potrebné analyzovať. Ak by sme chceli vedieť aktivitu v rámci IP adresy alebo podsiete, tak by prechádzanie všetkých záznamov predstavovalo značne časovo ale aj výpočetne náročnú úlohu, obzvlášť na veľkých chrbtových sietiach cez ktoré prechádza veľké množstvo dát. Preto je v našom záujme nájsť spôsob, pomocou ktorého by sme boli schopní efektívne prehľadávať dáta uložené v kolektore.

2. Analýza chovania IP adresy v čase

Cieľom tohto projektu je poskytnutie agregovaného pohľadu na koncové zariadenia v sieti, ktorý nám pomôže lokalizovať výkyvy v ich chovaní a identifikovať potenciálne hrozby. Agregovaným pohľadom rozumieme spojenie množstva údajov (napr. počet bajtov, paketov) za určité časové obdobie. Môžeme využiť vlastnosť kolektoru, ktorý surové záznamy tokov ukladá po určitých časových blokoch (napr. päťminútových intervaloch) a nad týmito blokmi vytvárať požadovaný pohľad.

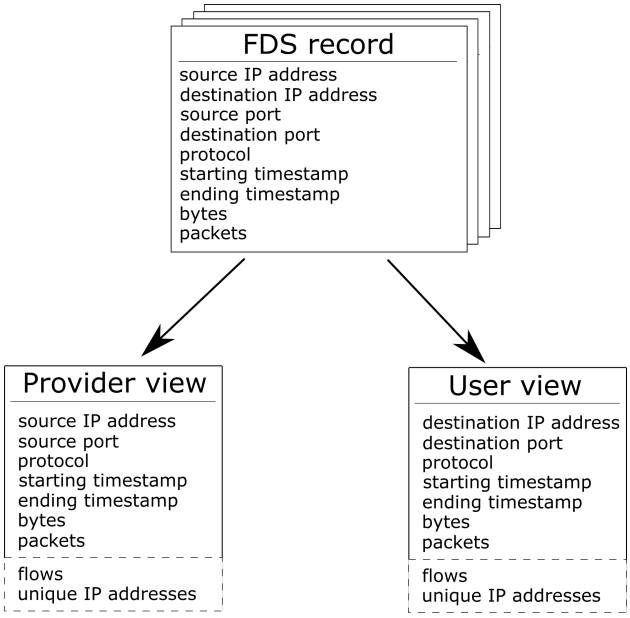
Základným riešením ktoré sa ponúka je extrahovanie záznamov z kolektora a ich uloženie vo forme binárneho súboru. Štruktúra súboru je prispôsobená zvýšeniu rýchlosti vyhľadávania a súbor je uložený pomocou kompresie pre minimalizovanie potrebného miesta na jeho uloženie. Cez veľké chrbtové siete prechádza veľké množstvo tokov, ktorých zdrojom ani cieľom nie sú zariadenia v monitorovanej sieti. Už v tomto bode sa nám ponúka možnosť urýchlenia budúceho prehľadávania a to v podobe filtrovania týchto tranzitných prenosov dát. Pri výbere tokov z kolektora si môžeme zvoliť konkrétne IP adresy alebo podsiete, ktoré sú pre nás zaujímavé, a tak sa zbaviť neužitočných tokov čo prispeje k urýchleniu vyhľadávania. Výsledný nástroj, ktorý je zobrazený na obrázku 2, vznikne spojením vyššie uvedených častí a bude realizovať filtráciu surových dát z kolektoru, agregáciu získaných štatistík a zápisu do binárneho súboru.



Obrázok 2. Nástroj FDS extractor

V prvom rade je potrebné definovať, čo budeme k jednotlivým koncovým zariadeniam reprezentovanými IP adresou zhromažďovať. Keďže nás zaujíma, aké služby dané zariadenie v čase využíva alebo poskytuje, budeme viazať jednotlivé štatistiky k L4 portom. Ku štatistikám, ktoré popisujú chovanie zariadenia a budú agregované, patria počet prenesených bajtov, paketov a všeobecne počet prenosov dát s inými zariadeniami. Aby sme vedeli rozoznať, aké množstvo

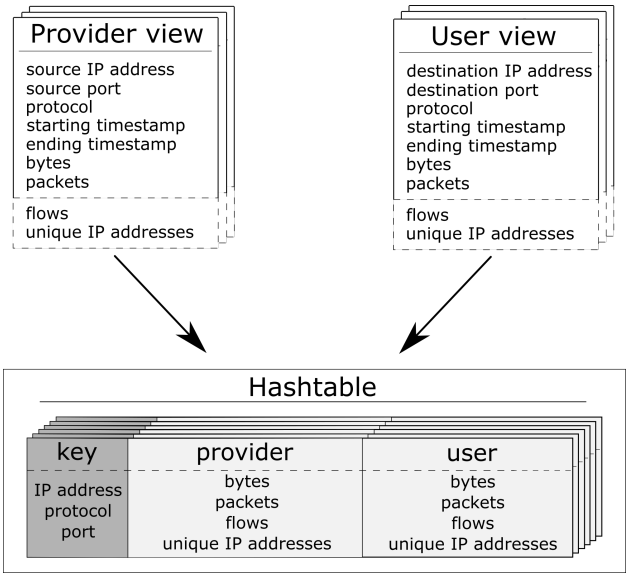
dát bolo poskytnutých konkrétnou IP adresou a aké množstvo adresa vyžadovala, rozdelíme si agregované dáta na 2 časti: spotrebované (user) a poskytnuté (provider). Medzi ďalšie informácie, ktoré môžu byť z nášho pohľadu zaujímavé sú napríklad počty adries, s ktorými IP adresa prišla do kontaktu alebo časové značky začiatku a konca komunikácie. Všetky tieto informácie prispievajú k určeniu základného chovania zariadenia a prípadne aj jeho typu. Tak si vieme zo záznamov vytvoriť pohľady na konkrétne chovanie IP adries, ako je ukázané na obrázku 3.



Obrázok 3. Pohľady na chovanie IP adresy získané zo záznamov prenosov

Samotný program, ktorý bude agregované štatistiky koncových zariadení vytvárať, bude pracovať so záznamami tokov uložených vo formáte súboru FDS. Na prácu s týmto súborom je použitá knižnica *libfds*[1], ktorá obsahuje funkcie a štruktúry určené pre prácu s týmto typom súboru. FDS súbor je formát súboru používaný open-source kolektorom IPFIXcol2 [5]. Samotný súbor obsahuje pôvodné NetFlow/IPFIX záznamy získané od sieťovej sondy a umožňuje s nimi pomocou uvedenej knižnice ďalej pracovať.

Záznamy NetFlow/IPFIX obsahujú získané informácie o tokoch v sieti. Konkrétne je z nich aplikáciou, ktorá vytvára pohľady, extrahovaná zdrojová a cieľová IP adresa, zdrojové a cieľové číslo portu, protokol a počet prenesených paketov a bajtov. Všetky tieto informácie sú dočasne vložené do tabuľky s rozptýlenými položkami, ktorá je zobrazená na obrázku 4, aby sme s nimi mohli manipulovať vrámci nášho nástroja. Kľúčom do tejto tabuľky je trojica (IP adresa, protokol, port), podľa ktorej je záznam indexovaný do tabuľky. Okrem týchto informácií sa pri nájdení toku navyšuje počítadlo nájdených záznamov v súbore a v prípade, že sa jedná o zariadenie, s ktorou doposiaľ IP adresa nekomunikovala, tak sa navýši aj počítadlo vzájomne komunikujúcich zariadení.



Obrázok 4. Tabuľka s rozptýlenými položkami obsahujúca pohľady s agregovanými štatistikami

Pred ukladaním extrahovaných údajov do tabuľky s rozptýlenými položkami je vhodné redukovať počet IP adries (alebo podsietí) na tie, ktoré nás zaujímajú (nie sú len tranzitné). Na to je využitý jednoduchý textový súbor obsahujúci zoznam IP adries zapísaných v bežnom prefixovom zápise. Z tohto súboru sú načítané IP adresy a podsiete, ktoré sa porovnávajú s aktuálne spracovávaným záznamom z FDS súboru. V prípade, že dôjde ku zhode IP adresy s nejakou IP adresou z filtru alebo IP adresa bude patriť do podsiete definovanej vo filtri, uloží sa IP adresa do tabuľky s rozptýlenými položkami.

[1] https://github.com/CESNET/libfds

3. Výstupný binárny súbor EXT1

Pred použitím binárneho súboru na uloženie extrahovaných dát je potrebné sa opýtať, či by nebolo efektívnejšie použiť jednu z konvenčných databáz (napr. *MySQL*, *InfluxDB*). Uvažovali sme aj nad použitím databáz ale nakoniec sme sa rozhodli pokračovať s binárnym súborom a to z dôvodu množstva štatistických záznamov, ktoré potrebujeme uložiť. Na veľkých chrbtových sietiach sa ich počet pohybuje v rádoch státisícov až miliónov. Pri exporte do bežných databáz je často nutné konvertovať ich obsah do textovej podoby a následne komunikovať s databázou. Pri takomto počte záznamov by to predstavovalo značnú réžiu, preto sme sa rozhodli využiť binárny súbor. Naviac sa samotné štatistiky vytvárajú nad FDS súbormi, ktoré obsahujú určitú množinu dát a je pre nás jednoduchšie vytvoriť také mapovanie, aby agregované štatistiky predstavovali pohľad nad dátami uloženými v súbore z ktorého vznikli.

Štruktúra súboru navrhnutého pre uloženie štatistických záznamov je zobrazená na obrázku 5 a skladá sa z troch základných častí:

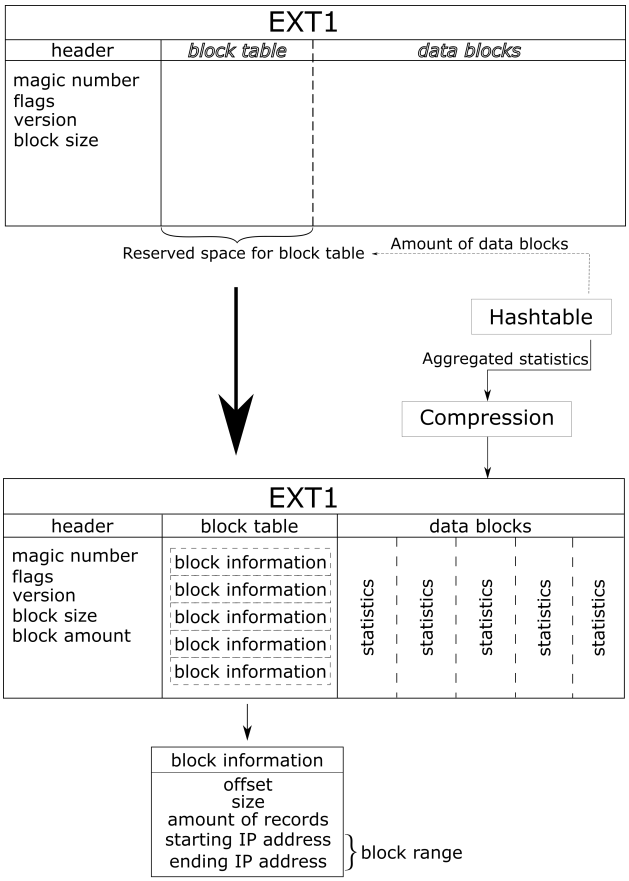
1. hlavička
2. tabuľka blokov
3. dátové bloky

Hlavička obsahuje informácie na identifikáciu súboru (magické číslo - EXT1), príznaky, verziu súboru, počet dátových blokov a veľkosť jedného dátového bloku pred použitím kompresie. Príznaky aktuálne obsahujú informáciu o type použitej kompresie.

Aby bolo možné v súbore rýchlo identifikovať pozíciu záznamov, ktoré nás zaujímajú, sú jednotlivé štatistiky IP adries zoradené podľa ich kľúča (trojica IP adresa, protokol, port) a uložené do dátových blokov. Vzhľadom k tomu, že pri vyhľadávaní nechceme strácať čas prechádzaním všetkých blokov, používa sa na začiatku umiestnená tabuľka blokov, ktorá jednotlivé bloky popisuje. Jej položky obsahujú informácie o začiatku jednotlivých blokov (relatívne posunutie od začiatku súboru - offset), veľkosti uloženého bloku po kompresii, počte záznamov v jednom bloku a rozsahu IP adries, ktoré daný blok pokrýva.

Postup uloženia agregovaných dát pre IP adresy je nasledovný. Po spracovaní všetkých záznamoch o tokoch zo vstupného súboru máme naplnenú tabuľku s rozptýlenými položkami so záznamami, ktoré potrebujeme uložiť. Na základe jej veľkosti je možné zistiť, koľko blokov bude potrebných na uloženie všetkých záznamov a rezervujeme si na začiatku súboru miesto potrebné pre hlavičku súboru a tabuľku blokov. Do hlavičky súboru zapíšeme informácie o súbore s počtom vypočítaných blokov. Postupne prechádzame tabuľku s rozptýlenými položkami. Vyberieme si vhodný počet záznamov, ktoré potrebujeme zapísať a pomocou kompresie ich zapíšeme na prvé voľné miesto v časti dátových blokov v súbore. Do tabuľky blokov vložíme informácie o zapísanom bloku a pokračujeme v spracovaní zvyšných záznamov. Ak už neostávajú žiadne záznamy na zápis, zapíšeme do súboru samotnú tabuľku blokov.

Opačný postup pri vyhľadaní konkrétnej IP adresy je nasledovný. Prechádzame položky tabuľky dátových blokov a kontrolujeme, či je IP adresa v rozsahu IP adries zapísaných v danom bloku. V prípade, že nájdeme správny blok, zistíme jeho relatívne posunutie oproti začiatku súboru a načítame príslušný dátový blok do pomocnej štruktúry (je rovnaká ako štruktúra bloku pre zápis). Pomocou dekompresie načítame údaje z dátového bloku a nájdeme požadovanú IP adresu. Týmto ušetríme množstvo prečítaných dát zo súboru čím zvýšime rýchlosť procesu vyhľadávania záznamu.



Obrázok 5. Štruktúra binárneho súboru EXT1

Jeden z hlavných problémov, ktorý som načrtol v tejto práci je množstvo uložených záznamov. Na veľkých chrbtových sietiach prechádza veľké množstvo dát, ktoré sú uložené v kolektore. Avšak aj tieto súbory obsahujúce agregované štatistiky zaberajú veľa miesta, čo predstavuje z dlhodobého hľadiska problém, pretože samotné záznamy je potrebné uchovávať po značne dlhú dobu až v radoch jednotiek mesiacov. Jedno z riešení je použitie kompresie pri zápise dátových blokov do súboru, čím sa zníži veľkosť vzniknutého binárneho súboru. Toto riešenie avšak predstavuje problém, pretože z hľadiska vyhľadávania využitie kompresie potenciálne negatívne ovplyvňuje rýchlosť vyhľadávania pretože pri čítaní je nutné použiť dekompresiu. Teda existuje tu nejaká miera kompromisu medzi rýchlosťou vyhľadávania a miestom potrebným na uloženie tokov zo siete.

Ďalší z problémov spočíva v tom, ako efektívne zistiť počet unikátnych IP adries (obrázok 4), s ktorými sledovaná IP adresa komunikovala. Táto informácia nám môže slúžiť napríklad na odhalenie zariadenia, ktoré skenuje IP adresy v sieti, pretože odpovedajúci počet adries pre dané zariadenie bude viacnásobne vyšší ako pre bežné zariadenia. Keďže potrebujeme zistiť počet unikátnych IP adries, budeme si priebežne ukladať IP adresy, s ktorými zariadenie doposiaľ nekomunikovalo. Po prejdení všetkých vstupných dát budeme mať k dispozícii presný počet unikátnych IP adries. V prípade tohto postupu môže nastať problém, ak by došlo k masívnemu skenovaniu siete, ktoré by spôsobilo významné spomalenie výpočtu a zvýšenie použitia pamäte. Alternatívny spôsob spočíva v aproximácií počtu adries a dá sa aplikovať pomocou bloomovho filtra [2]. Bloomov filter sa používa na overenie príslušnosti prvku do množiny. Na odhadnutie výsledku využíva pravdepodobnosť, preto výsledok môže byť chybný (čo nám v našej aplikácií neprekáža z dôvodu spomenutého vyššie). Jeho výhodou je časová a priestorová efektivita. Ku nevýhodám patrí nutnosť správne odhadnúť jeho veľkosť a s tým spojený počet hašovacích funkcií a možné kolízie kľúčov.

4. Plány na pokračovanie v projekte

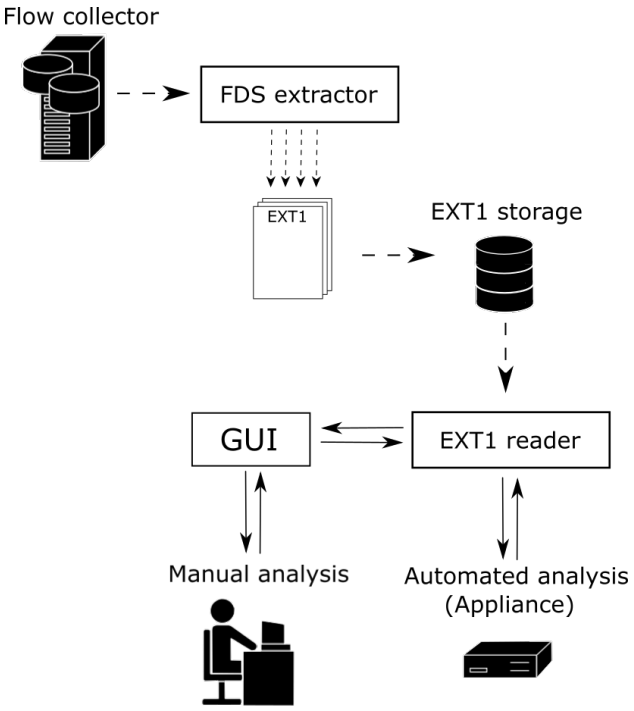
Aktuálne je v projekte kompresia zabezpečená pomocou algoritmu *zstd*. Jeden z plánov pre pokračovanie projektu je implementovať kompresiu pomocou ďalších algoritmov (napr. *LZ4*) a porovnať ich výsledky v rámci rýchlosti vyhľadávania informácií o komunikácií IP adries a veľkosti vytvoreného binárneho súboru.

Ďalší potenciálny smer by mohol spočívať v zmene štruktúry, do ktorej sa extrahujú záznamy zo vstupného FDS súboru. V súčasnosti je použitá tabuľka s rozptýlenými položkami a poradie je stanovené reláciou usporiadania nad jej kľúčmi. Je veľa faktorov, ktoré ovplyvňujú rýchlosť zoradenia položiek v dátovej štruktúre, napríklad počet vstupných údajov, vybraný algoritmus na zoradenie alebo typ samotnej štruktúry na uloženie údajov. Za vyskúšanie by stálo použiť viacero možných kombinácií štruktúr a algoritmov na radenie a porovnávať ich namerané rýchlosti.

Výpočet unikátnych IP adries v súčastnosti nie je implementovaný v programovom riešení projektu. Bolo by vhodné ho tam pri pokračovaní práce doplniť spolu s vyššie uvedením bloomovým filtrom, keďže by sme získali pomerne zaujímavú štatistiku popisujúcu chovanie zariadenia.

Vytvorenie binárneho súboru s agregovanými štatistikami predstavuje iba prvú časť procesu urýchlenia analýzy chovania koncových staníc. Druhou potrebnou časťou je nástroj, ktorý by nám umožnil čítať a interpretovať získané štatistiky. Otázne je, akým spôsobom by tieto nástroje mali spolupracovať. Jednou z možností je extrahovanie agregovaných štatistík a ich následná interpretácia len pre niektoré súbory z dátového kolektoru. Iba z týchto súborov, ktoré by obsahovali informácie o tokoch z analyzovaného časového obdobia, by boli vytvorené agregované štatistiky vo forme súborov EXT1, ktoré by boli následne interpretované. V tomto prípade by však do času potrebného na interpretáciu zasahoval aj čas potrebný na extrahovanie štatistík a analyzovanie dlhšieho časového obdobia by bolo časovo náročné. Prijateľnejším riešením je priebežné vytváranie agregovaných pohľadov na dáta priamo z kolektoru, ktoré je zobrazené na obrázku 6. Binárne súbory EXT1 obsahujúce tieto štatistiky budú takýmto spôsobom okamžite k dispozícii na prípadné čítanie. Na prvý pohľad sa môže zdať, že prakticky zdvojíme už raz namerané dáta, čo ale nie je pravda, pretože vytvárame len agregované pohľady určitých podsietí, ktoré sú navyše komprimované, a oproti pôvodným záznamom vyžadujú menej miesta na ich uloženie.

Ďalším rozšírením by bola interpretácia agregovaných štatistík. Pre potreby manuálnej analýzy sa ponúka možnosť vytvorenia užívateľského rozhrania, ktoré by tieto štatistiky zobrazovalo.



Obrázok 6. Nadväzovanie čítania na zápis agregovaných štatistík

5. Záver

Podstatou tejto práce bolo extrahovanie informácií získaných monitorovaním siete a vytvorenie agregovaného pohľadu na komunikáciu koncových staníc v sieti takým spôsobom, aby bolo možné v prípade výskytu bezpečnostného incidentu rýchlo lokalizovať bod, v ktorom sa incident prejavil.

V teoretickom úvode práce bolo predstavené monitorovanie siete, aká za ním existuje motivácia a aké sú základné spôsoby jeho realizácie. Taktiež sme načrtli problém spojený s množstvom dát, ktoré je potrebné počas analýzy bezpečnostného incidentu spracovať.

Na začiatku hlavnej časti sme predstavili základné riešenie uvedeného problému. Toto riešenie je podrobne popísané v ďalších častiach tejto práce. Hlavnou myšlienkou riešenia je vytvorenie agregovaného pohľadu na jednotlivé koncové stanice a jeho uloženie do výstupného binárneho súboru, ktorého štruktúra je prispôsobená rýchlemu vyhľadávaniu.

Ďalej sme predstavili, aké informácie o koncových zariadeniach je potrebné extrahovať a kým spôsobom ich môžeme získať z dátového kolektoru. Popísali sme dátovú štruktúru použitú na uloženie štatistík o komunikácií zariadení a filtrovanie tranzitných prenosov.

Diskutovali sme použitie konvenčných databáz a dospeli sme k záveru, že použitie binárneho súboru je lepším spôsobom uloženia získaných informácií. Vytvorili sme vlastný binárny súbor EXT1, ktorého štruktúra je prispôsobená rýchlemu vyhľadávaniu potrebných záznamov. Popísali sme jeho štruktúru, spôsob akým sa do neho ukladajú získané záznamy a akým spôsobom by bolo možné realizovať čítanie zapísaných dát.

Na záver sme predstavili možné rozšírenia tejto práce, ktoré spočívajú v experimentovaní s použitou kompresiou, dátovou štruktúrou na uloženie extrahovaných záznamov, doplnení štatistiky unikátnych IP adries a vytvorení komplementárneho nástroju, ktorý by nám umožňoval čítať a interpretovať pohľady z binárneho súboru EXT1.

Poďakovanie

Rád by som poďakoval Ing. Lukášovi Hutákovi za odbornú pomoc a Ing. Martinovi Žádníkovi Ph.D., ktorý mi dal možnosť zúčastniť sa tohto projektu.

Literatúra

[1] CLAISE, B., TRAMMELL, B. a AITKEN, B. Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information. september 2013. Dostupné z: https://www.rfc-editor.org/info/rfc7011.

[2] CONTRIBUTORS, W. *Bloom filter — Wikipedia, The Free Encyclopedia*. 2022. [Online; navštívené 26-Január-2022]. Dostupné z: https://en.wikipedia.org/wiki/Bloom_filter.

[3] CONTRIBUTORS, W. *NetFlow — Wikipedia, The Free Encyclopedia*. 2022. [Online; navštívené 26-Január-2022]. Dostupné z: https://en.wikipedia.org/wiki/NetFlow.

[4] HOFSTEDE, R., CELEDA, P., TRAMMELL, B., DRAGO, I., SADRE, R. et al. Flow Monitoring Explained: From Packet Capture to Data Analysis With NetFlow and IPFIX. 2014, roč. 16, s. 2037–2064. ISSN 1553-877X.

[5] HUTÁK, L. Nová Generace IPFIX Kolektoru. 2018. Dostupné z: https://dspace.vutbr.cz/bitstream/handle/11012/84955/final-thesis.pdf.