UNIVERZITA PAVLA JOZEFA ŠAFÁRIKA V KOŠICIACH
PRÍRODOVEDECKÁ FAKULTA

ČASOPRIESTOROVÁ ANALÝZA V KYBERNETICKEJ BEZPEČNOSTI

2024 **Filip DVORSKÝ**

UNIVERZITA PAVLA JOZEFA ŠAFÁRIKA V KOŠICIACH
PRÍRODOVEDECKÁ FAKULTA

ČASOPRIESTOROVÁ ANALÝZA V KYBERNETICKEJ BEZPEČNOSTI

BAKALÁRSKA PRÁCA

Študijný program:	informatika
Pracovisko (katedra/ústav):	Ústav informatiky
Vedúci diplomovej práce:	doc. RNDr. JUDr. Pavol Sokol, PhD.

Košice 2024

Filip DVORSKÝ

Univerzita P. J. Šafárika v Košiciach
Prírodovedecká fakulta

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Filip Dvorský
Študijný program: informatika (jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: Informatika
Typ záverečnej práce: Bakalárska práca
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Časopriestorová analýza v kybernetickej bezpečnosti

Názov EN: Spatio-temporal analysis in cyber security

Cieľ: (1) Úprava dátovej sady pre časopriestorovú analýzu dát v kybernetickej bezpečnosti.
(2) Porovnanie aktuálnych prístupov k časopriestorovej analýze v kybernetickej bezpečnosti.
(3) Analýza vybraných prístupov nad vytvorenou dátovou sadou, vyhodnotenie a interpretácia výsledkov.

Literatúra: (1) Amin, R. W., Sevil, H. E., Kocak, S., Francia III, G., & Hoover, P. (2020). The spatial analysis of the malicious uniform resource locators (URLs): 2016 dataset case study. Information, 12(1), 2.
(2) Sethi, A. A. R. U. S. H. I. "Statistical Methods for Spatial Data Analysis." Cyber Secur. Insights Mag 1 (2022): 7-11.
(3) Sokol, Pavol, and Veronika Kopčová. "Lessons learned from correlation of honeypots' data and spatial data." 2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE, 2016.

Vedúci: doc. RNDr. JUDr. Pavol Sokol, PhD.
Oponent: RNDr. Tomáš Bajtoš
Ústav : ÚINF - Ústav informatiky
Riaditeľ ústavu: doc. RNDr. Ondrej Krídlo, PhD.

Dátum schválenia: 14.05.2024

Poďakovanie

Týmto sa chcem poďakovať vedúcemu práce doc. RNDr. JUDr. Pavlovi Sokolovi, PhD. za vecnú a cennú pomoc pri výskume a písaní, za odborné vedenie, užitočné návrhy a za cielene smerovanie počas tvorby práce.

Abstrakt v štátnom jazyku

Každý útok v kybernetickom priestore sa uskutočňuje v určitom čase a na konkrétnom mieste. Okrem štandardného času cieľa útoku je možné vziať do úvahy aj čas zdroja útoku. Chápanie času a priestoru útokov nám umožňuje lepšie pochopiť spôsob uskutočnenia útokov zo strany útočníkov. To nám umožňuje efektívnejšie nastaviť bezpečnostné politiky a postupy v organizácii za účelom adekvátnej reakcie na bezpečnostné útoky a incidenty. V práci sa venujeme návrhu a implementácii časopriestorovej analýzy v kybernetickej bezpečnosti. Ako zdroje údajov používame záznamy bezpečnostných udalostí zo systému Warden. V práci sa zameriavame na úpravu týchto údajov vrátane ich obohatenia z externých zdrojov. Obohatenie sa dotýka najmä bližšieho popisu vrátane reputácie IP adresy. V práci sa zameriavame na analyzovanie jednotlivých bezpečnostných udalostí z pohľadu zdroja útoku, a teda na umiestnenie a lokálny čas daného zdroja útoku. Pre účely časopriestorovej analýzy sme získané bezpečnostné udalosti vyjadrili v podobe časových radov a tepelných máp.

Kľúčové slová: bezpečnostné udalosti, IP adresa, časopriestorová analýza, časové rady, kybernetická bezpečnosť

Abstrakt v cudzom jazyku

Every attack in cyberspace occurs at a certain time and place. In addition to the standard time of the attack target, it is also possible to consider the time of the attack source. Understanding the time and space of attacks allows us to better understand the methods used by attackers. This enables us to set security policies and procedures more effectively within an organization to adequately respond to security attacks and incidents. In this work, we focus on the design and implementation of spatiotemporal analysis in cybersecurity. We use security event records from the Warden system as our data sources. Our work involves modifying these data, including enriching them from external sources, particularly focusing on detailed descriptions such as IP address reputation. We analyze individual security events from the perspective of the attack source, focusing on the location and local time of the attack source. For the purpose of spatiotemporal analysis, we represent the obtained security events in the form of time series and heatmaps.

Keywords: security events, IP address, spatiotemporal analysis, time series, cybersecurity

Obsah

Zoznam ilustrácií

Úvod

Rastúca zložitosť informačných technológií so sebou prináša rast v podobe počtu bezpečnostných hrozieb a následne bezpečnostných incidentov. Najmä aj vďaka rastúcej zložitosti je pre organizácie čoraz ťažšie zabezpečiť ochranu informačných systémov a technológií. Preto môžeme vidieť rastúci záujem organizácii ohľadom zabezpečenia ich kritickej infraštruktúry a kybernetickej bezpečnosti ako takej. Tento trend organizácii sa teda jasne pretavuje v podobe investícii do technológii a postupov, ktoré umožňujú prevenciu a predikciu daných kybernetických útokov na ich infraštruktúru.

V istých prípadoch samotná prevencia ako forma pasívnej ochrany nie je dostačujúca. Preto musíme siahnuť po efektívnejších nástrojoch pre pochopenie význačnosti jednotlivých kybernetických útokov. Pre tieto prípady vieme použiť časopriestorovú analýzu, vďaka ktorej by sme mali lepšie predikovať dané kybernetické útoky a tým aj lepšie nastaviť naše vnútorne politiky.

Medzi hlavne ciele tejto prace patri vykonanie samotnej časopriestorovej analýzy nad upravenou dátovou sadou spolu s interpretáciou výsledkov pomocou časových radov a tepelných máp. Pričom zameranie nasej časopriestorovej analýzy stoji na čase kedy zneužité zdrojové zariadenie útoku začalo vykonávať daný útok.

Prvým cieľom našej práce patrí spracovanie, obohatenie a upravenie dátovej sady do formy, pomocou ktorej môžeme vykonať časopriestorovú analýzu. Ako druhý cieľ máme porovnať a opísať jednotlivé, už používané, prístupy k analýze dát pomocou časopriestorovej analýzy. V poslednom cieli mame analyzovať jednotlivé vybrané prístupy nad našou dátovou sadou spolu s interpretáciou výsledkov našich analýz. Pričom naše analýzy sú zamerane na čas zdrojového zariadenia útoku.

Našu prácu sme rozdelili do piatich hlavných kapitol. V prvej kapitole vysvetľujeme zakladane pojmy a princípy, najmä priestorové a časove údaje, ktoré majú vplyv na pochopenie a interpretáciu našej dátovej sady spolu s vysvetlením bezpečnostných udalostí. V druhej kapitole sa zameriavame na našu dátovú sadu spolu s jej rôznymi formátmi. Definujeme, z akých zariadení a systémov pochádza naša dátová sada, a teda aj akú ma štruktúru. V tretej kapitole vysvetľujeme geolokačné a reputačné služby ako také. Súčasne analyzujeme ich vplyv a dôležitosť na našu časopriestorovú analýzu spolu s vysvetlením dôvodov výberu jednotlivých geolokačných a reputačných služieb. V štvrtej kapitole opisujeme samotnú implantáciu spracovania, obohatenia

a upravenia našej dátovej sady do formy, s ktorou môžeme vykonať našu časopriestorovú analýzu. V piatej kapitole vysvetľujeme princípy a realizáciu časopriestorovej analýzy nad našou dátovou sadou spolu s interpretáciou výsledkov pomocou časových radov a tepelných máp.

1 Priestorové a časové údaje

Na pochopenie významu časopriestorovej analýzy je potrebné, aby sme najprv vysvetlili, čo sú to časové a priestorové údaje. V nasledujúcich častiach vysvetlíme význam týchto údajov pre našu analýzu a napokon ako sú reprezentované v našej dátovej sade.

1.1 Priestorové údaje

Priestorové údaje reprezentujú bodové alebo plošne objekty, ktoré sa nachádzajú v geografickom priestore. Priestorové údaje zahŕňajú informácie o fyzickej polohe, tvare a vlastnostiach objektov. Tieto údaje môžu byt reprezentovane ako zemepisná šírka a dĺžka, adresy a mnoho ďalších [1].

Z hľadiska našej práce sa budeme hlavne venovať geolokácii IP adresy. V práci budeme predpokladať s veľmi vysokou pravdepodobnosťou, že IP adresy, nachádzajúce sa v našej dátovej sade a priradené k zdrojom útokov, sú priradené k zariadeniam, ktoré boli zneužité útočníkom na nejaký útok. Súčasne budeme predpokladať, že k daným zariadeniam je možne priradiť nejakú lokalitu, resp. iný priestorový údaj. Inak povedané, zdroj útoku (zneužité zariadenie) vieme reprezentovať pomocou približnej zemepisnej šírky a zemepisnej dĺžky, prípadne všeobecnejšie podľa krajiny alebo časového pásma, v ktorom sa dané zariadenia nachádza.

1.2 Časové údaje

Časove údaje reprezentujú všetky informácie, ktoré sú spojene s časom. Môže ísť o merania, udalosti alebo pozorovania, ktoré boli zachytené v nejakých časových bodoch, ako napríklad úder blesku a podobne [2].

Z hľadiska našej práce budeme za jeden z hlavných časových údajov považovať čas detekcie. Tento čas reprezentuje moment v čase, kedy bola bezpečnostná udalosť detegovaná na nejakom detekčnom zariadení. Lokálny čas zdroja útoku (zneužitého zariadenia) je časový udaj, ktorý predstavuje ten istý moment v čase ako čas detekcie. Avšak jeho reprezentácia je závislá od časovej zóny, v ktorej sa zneužité zariadenie nachádza. V našej práci sa budeme pozerať na to, ako lokálny čas zdroja útoku (zneužitého zariadenia) vplýva na frekvenciu, resp. počet útokov na organizáciu.

1.3 Bezpečnostná udalosť

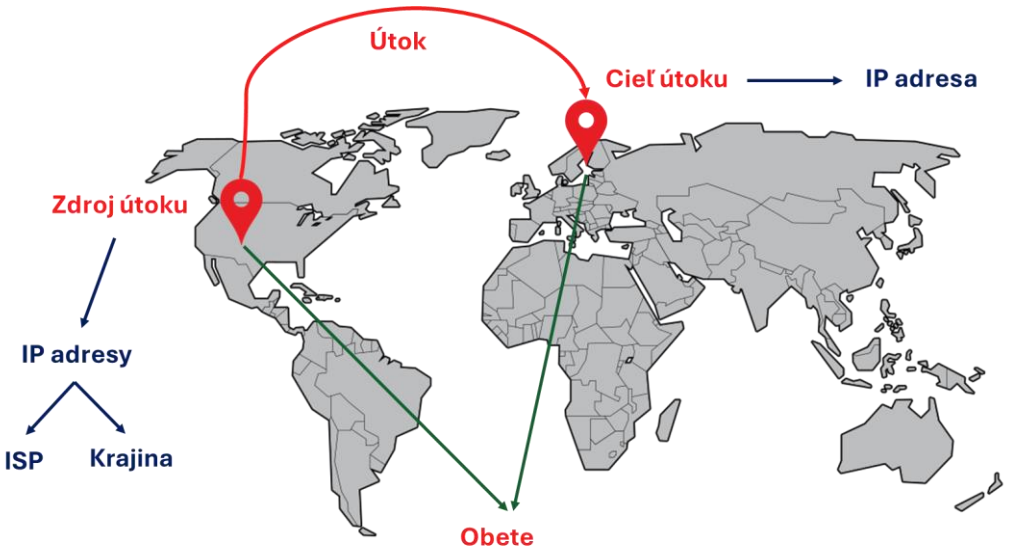
Udalosť v pojímaní informačných a komunikačných technológii môžeme považovať ako nejaký pozorovateľný jav, ku ktorému došlo v určitom časovom bode v nejakom systéme alebo v sieti [3].

Pod pojmom **bezpečnostná udalosť** budeme považovať udalosť pre, ktorú platí, že môže mat potenciálny vplyv na dôvernosť, integritu a dostupnosť nejakého informačného systému [4].

Je doležíte podotknúť význam slova potenciálny, keďže v našej práci sa budeme venovať bezpečnostným udalostiam z rôznych detekčných systémov. Z toho vyplýva, že rôzne detekčne systémy môžu považovať nejakú aktivitu za bezpečnostnú udalosť pričom iné detekčné zariadenie by takúto aktivitu vyhodnotilo ako štandardnú. Tieto rozdiely sú závislé od zariadenia k zariadeniu, pričom väčšina detekčných zariadení je konštruovaná na detekciu špecifických anomálii.

1.4 Schéma útoku

Keďže bezpečnostné udalosti sú priamo informácie o už vykonanom útoku, je doležíte si vysvetliť, ako samotný útok vyzerá. To je dôležité pre lepšie pochopenie údajov uložených v bezpečnostnej udalosti.



Obrázok 1: Schéma útoku

Ako môžeme vidieť z Obr. č. 1, útok ma stále nejaký zdroj a ciel. Je doležíte podotknúť, že v oboch prípadoch sa jedna o obeť, keďže útočníci na útoky zneužívajú kompromitované zariadenia. Následne zo zdroja vieme získať IP adresu, pomocou ktorej môžeme následne cez externé služby získať informácie o krajine, poskytovateľovi internetovej služby, približnú lokáciu a mnoho ďalších. A podobne ako pri zdroji útoku aj pri cieli vieme získať podobné údaje.

1.5 Podobné práce

Problematike, ktorú v tejto bakalárskej práce analyzujeme, sa venuje viacero výskumných skupín. Nami prezentovaná práca vychádza z týchto prác. V tejto kapitole si uvedieme niekoľko prác relevantných nášmu výskumu s uvedením spoločných a rozdielnych čŕt medzi našimi výskumnými aktivitami.

Tabuľka 1: Porovnanie podobných prác

Článok	Typ dát	Metóda	Typ analýzy	Cieľ
[5]	Dátová sada malicióznych URL kategorizovaných podľa typu útoku.	Kartogramové mapy	Priestorová	**Demonštrácia vizuálnych analýz a ich prínos pre bezpečnostného analytika**
[6]	Dátová sada príjmov občanov pre vybrané krajiny	Štatistická analýza	Priestorová	**Preukázanie významnosti štatistickej analýzy pre priestorové dáta**
[7]	Dátová sada bezpečnostných udalosti z detekčných zariadení typu honeypot.	Korelačná analýza	Priestorová	**Ukážka korelačnej analýzy pre priestorové dáta a dáta z honeypotov**
[8]	Dátová sada z detekčných zariadeniu typu honeypot.	Tepelné mapy	Časová	**Demonštrácia práce z časovo orientovanými dátami a interpretácia výsledkov analýz pomocou tepelných máp**
[9]	Dátová sada z detekčných zariadení typov	Kauzálna analýza	Priestorová analýza	**Ukážka kauzálnej analýzy útokov nad dátovou sadou**

	honeypot spolu z dátami o ekonomických a sociálnych vlastnostiach krajín			**z honepotov založenej na vlastnostiach krajín.**
[10]	Dátová sada správ bezpečnostných incidentov	Blacklisting, Štatistická analýza	Reputačná analýza	**Demonštrovanie výhod využitia analýz pre optimalizáciu kategorizácie malicióznych IP adries**
[11]	Dátová sada bezpečnostných údajov z rôznych externých služieb, stránok a otvorených zdrojov	Geolokácia IP adries	-	**Návrch a implementácia nástroja na spracovanie dát pre bezpečnostného analytika z otvorených zdrojov.**
[12]	Dátová sada bezpečnostných údajov a scenárov z otvorených zdrojov.	Predikčne metódy	Časová a priestorová analýza	**Návrh a demonštrácia predikčných metód postavených na časovej a priestorovej analýze**

Autori Amin a spol. sa v práci [5] venujú analýze a vizualizácii škodlivých webových stránok, následnej klasifikácii do jednotlivých typov ako napríklad phishing či spam. Autori následne vizualizujú výslednú dátovú sadu do rôznych geograficky máp podľa krajiny a iných parametrov.

Autor sa v práci [6] zmeral na popis štatistických metód možných pre využitie v analýze priestorových dát. Súčasne popísali aj možné techniky vizualizácie v podobe kartogram máp (choropletová mapa) spolu s princípmi a radami uplatniteľnými pre štatistickú analýzu dát.

Sokol a spol. sa v článku [7] venovali problematike zdrojov útokov. Ako dátovú sadu použili údaje z honeypotov a honeynetov a nad danou sadou vykonali v rámci práce aplikáciu tepelných máp (heatmapy). Títo autori popísali vplyv počtu útokov a vlastnosti krajín, z ktorých útočníci útočia, na analýzu dátovej sady.

Autori Sokol a spol. sa v článku [8] venujú spracovaniu dát z honeypotov a honeynetov za účelom analýzy a vizualizácie primárne časových dát. Popisujú nastavenie procesu analýzy pre už spomínanú časovú analýzu. Taktiež sa venujú implementácii vizualizácie tejto analýzy v podobe tepelných máp (heatmaps).

Autori Zuzčák a Bujok sa v práci [9] venujú analýze bezpečnostných dát z pohľadu štatistickej analýzy vzhľadom na jednotlivé charakteristiky krajín, v ktorých sa nachádzajú zdrojové zariadenia útokov. Následne sa autori venujú analýze socioekonomických vlastnosti jednotlivých krajín a ich vplyvu na počet bezpečnostných udalostí.

Autori Carriegos a spol. sa článku [10] venujú analýze reputácie IP adries a ich následnej analýze vzťahov medzi jednotlivými reputačnými databázami. Autori sa následne venujú interpretácii výsledkov a predikcii s využitím matematických metód.

Autor Čergeť a spol. sa v práci [11] venujú problematike geolokácii IP adresy z externých služieb a otvorených zdrojov z cieľom vytvorenia nastroja na analýzu bezpečnostných udalosti z pohľadu zdroja útoku. Ďalej sa autori venujú rôznym porovnaniam ich nastroja z reálnymi dátami o zdrojových zariadenia za účelom zjednodušenia práce bezpečnostného analytika.

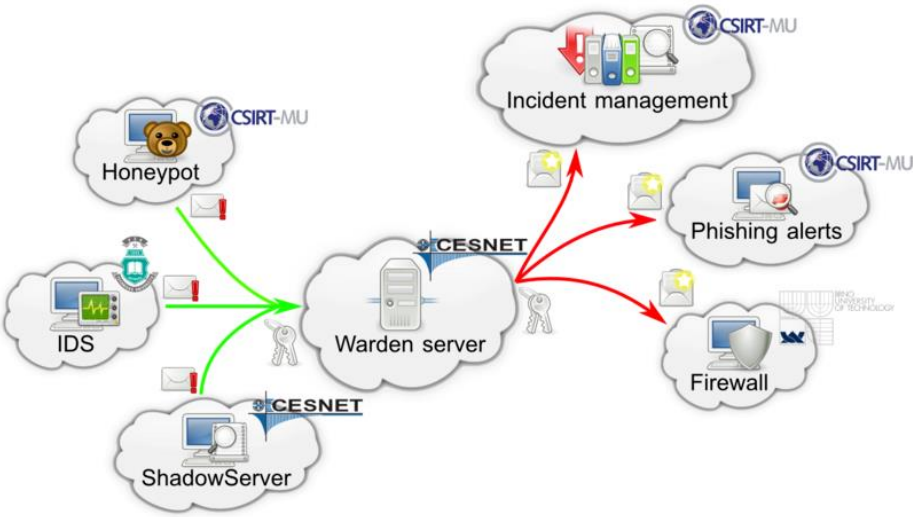
Autori Tan a spol. sa v článku [12] venujú popísaniu a implementácii predikčnej metódy založenej na časo-priestorovej analýze. Následne sa autori venujú komparačnej analýze ich metódy pomocou verejnej DARPA2000 dátovej sady scenárov, porovnávajú ich novo vytvorenú metódu oproti štandardne používaným metódam stanoveným v článku.

V našej práci sa venujeme implementácii časopriestorovej analýzy. S vyššie uvedenými prácami spolu máme podobný prístup k analýzam a vizualizácii výsledkov pomocou tepelných máp a časových radov. Väčšina prác sa zameriava najmä na lokáciu, odkiaľ prichádza útok. Časový aspekt v podobne uplatnenia časovej zóny je minoritná. Súčasne sa naša práca od predchádzajúcich výskumných prác odlišuje tým, že sa primárne zameriavame na zariadenie zneužité útočníkom. Na tieto zariadenia pozeráme z pohľadu času a priestoru, pričom zohľadňujeme ďalšie atribúty (ako je napr. reputácia).

2 Dátová sada

Všetky dáta v našej dátovej sade sú zo systému Warden [13], ktorý bol vytvorený a je prevádzkovaný spoločnosťou CESNET. Táto spoločnosť združuje vysoké školy a Akadémiu vied Českej republiky. Systém Warden je jeden z produktov vyvinutých spoločnosťou CESNET pre zefektívnenie spravovanej infraštruktúry.

Veľká časť udalostí v našej dátovej sade pochádza pravé z pascí na útočníkov, ktoré označujeme ako honeypoty. **Honeypoty** predstavujú zariadenia, ktoré slúži ako návnada, alebo klamný ciel pre útočníkov z cieľom ich nalákať a pozorovať/odvrátiť útoky [14].



Obrázok 2: Schéma Warden systému. Prevzaté z [13]

Na Obr. 2 môžeme vidieť zjednodušenú schému systému Warden, v rámci ktorej Warden server predstavuje centrálny server, ktorý spracováva bezpečnostne upozornenia od rôznych typov zariadení (honeypoty, detekčné zariadenia,…) a posiela dané upozornenia vybraným systémom vo forme IDEA štruktúrovaných udalosti.

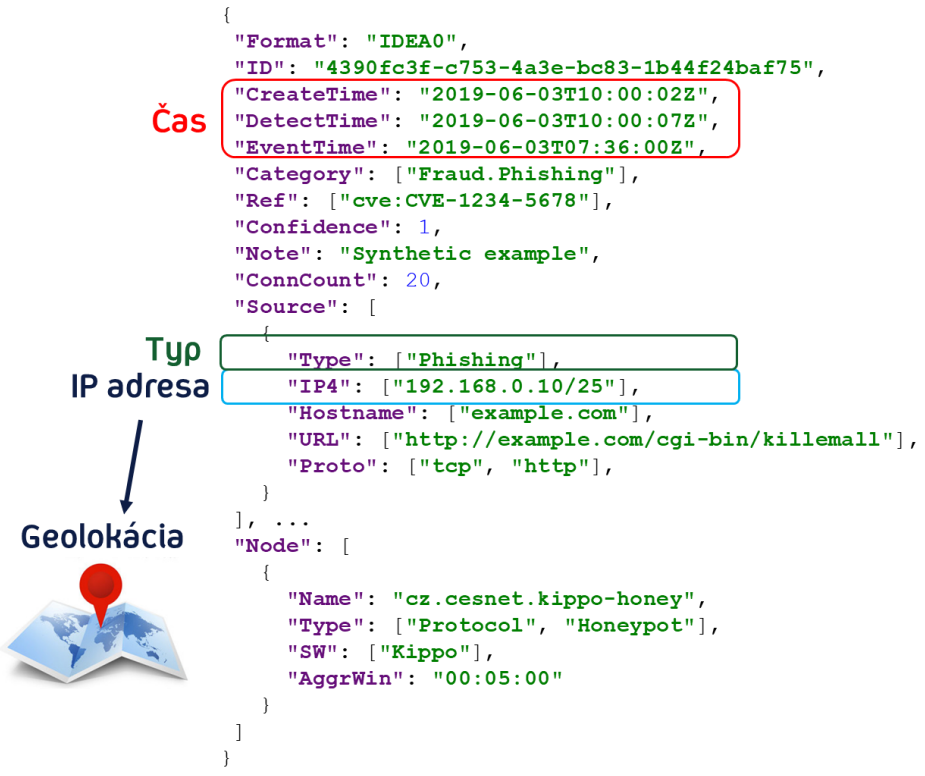
2.1 IDEA štruktúrovaný formát

V rámci systému Warden sa získané údaje ukladajú vo formáte IDEA (Intrusion Detection Extensible Alert). Bezpečnostne udalosti používane v IDEA rámci sú postavene na formáte JSON, a teda pre naše potreby budeme musieť dane udalosti upraviť do formátu nám lepšie spracovateľnému.

```
{"DetectTime": "2021-12-04 00:02:18+01:00", "Category": ["Attempt.Login",
"Test"], "Description": "SSH dictionary/bruteforce attack", "Format": "IDEA0",
 "CreateTime": "2021-12-04 00:02:19+01:00", "Note": "Banned by fail2ban",
"Source": [{"IP4": ["59.29.227.55"], "Proto": ["tcp", "ssh"]}], "ID":
"c0c764c3-fd00-499e-8467-a9b086850dea", "Node": [{"Note": "UPJS warden_filer
receiver", "SW": ["warden_filer-receiver"], "Type": ["Relay"], "Name":
"sk.upjs.science.warden_reader"}, {"Type": ["Relay"], "Name":
"cz.cesnet.syslog.warden_filer"}, {"SW": ["Fail2Ban"], "Type": ["Log",
"Statistical"], "Name": "cz.cesnet.syslog.sshd_block"}], "Target": [{
"Hostname": ["luna70.fzu.cz"], "Port": [22], "Proto": ["tcp", "ssh"]}]}
```


Obrázok 3: Ukážka bezpečnostnej udalosti v IDEA formáte

IDEA formát predstavuje flexibilný a praktický formát pre komunikáciu bezpečnostných udalostí medzi honeypotmi a inými detekčnými zariadeniami [15]. Pričom rozhodnutie postaviť IDEA na textovom formáte JSON zabezpečuje stabilnú spätnú kompatibilitu so staršími zariadenia, ale aj možnosť využiť zariadenia s menším výpočtovým výkonom.



Obrázok 4: Lokácie nutný parametrov časopriestorovej analýzy v IDEA formáte.

Obr. č. 4 je ukážkou syntetickej bezpečnostnej udalosti v IDEA formáte, spolu s najdôležitejšími údajmi pre vykonanie časopriestorovej analýzy, ako napríklad čas a IP adresa.

2.2 Kategórie bezpečnostných udalostí v dátovej sade

Z hľadiska architektúry Warden sa pomocou detekčných zariadení nachádzajúcich sa v nej vytvárajú bezpečnostné udalosti spadajúce do rôznych kategórií. Niektoré kategórie bezpečnostných udalosti sú všeobecné v ponímaní veci, ktoré spadajú do určitých kategórií. Napríklad medzi tieto kategórie patria Recon.Scanning a Attempt.Login, ktoré sú aj najpočetnejšie v našej dátovej sade. Iné kategórie sú viac špecifickejšie, ako napríklad Intrusion.AdminCompromise. Celkovo sa v našej dátovej sade nachádza 24 rôznych kategórii bezpečnostných udalostí.

Tabuľka 2: Prvých 5 najpočetnejších kategórií bezpečnostných udalosti

Kategória	Počet
Test	269 351 978
Attempt.Login	267 678 468
Recon.Scanning	215 277 252
Intrusion.UserCompromise	6 464 171
Availability.DoS	3 256 872
Attempt.Exploit	1 980 300

Z Tab, č. 1 je vidieť, že najpočetnejšia kategória Recon.Scanning tvorí až 94 % všetkých bezpečnostných udalostí. Bezpečnostná udalosť typu Recon.Scanning je udalosť, pri ktorej bolo detegované, že daný útočník sa snažil zistiť, aké zariadenie, typ zariadenia, otvorené porty alebo aj aké služby sa na detekčnom zariadení nachádzajú. Medzi druhé najpočetnejšie sa radi Attempt.Login s počtom 5,3 milióna bezpečnostných udalostí, tvoriac 3% všetkých bezpečnostných udalostí. Attempt.Login je kategória, v rámci ktorej sa útočník snažil prihlásiť do daného detekčného zariadenia, či už ako

používateľ alebo ako administrátor. Je dôležité podotknúť, že prihlásenie nemusí byť úspešne na to, aby dané detekčne zariadenie vyhodnotilo túto aktivitu ako bezpečnostnú udalosť.

2.3 Štruktúra dátovej sady

Naša dátová sada je zložená z 364 samostatných SQLite databázových súborov. Celková veľkosť našej dátovej sady je približne 681 GB. Ide celkovo o 500.825.074 bezpečnostných udalostí rôznych typov. Celá dátová sada je za obdobie od januára 2021 do decembra 2021 s výnimkou mesiaca júl 2021. Súčasne sme k dátovej sade pridali bezpečnostné udalosti za mesiac apríl 2024. To bolo najmä z dôvodu aplikácie navrhnutých analýz aj nad aktuálnymi údajmi.

	data	time	id
	Filter	Filter	Filter
1	{"DetectTime": "2021-12-04 ...	2021-12-04T00:00:01Z	fd56d4c8-ec65-4c76-8c34-0388334856f3
2	{"DetectTime": "2021-12-04 ...	2021-12-04T00:00:05Z	916bbb4b-c975-4770-8405-8ecaff4b6787
3	{"DetectTime": "2021-12-04 ...	2021-12-04T00:00:33Z	54f602fb-f362-4ac4-bf2f-a5da8bcbbb8a
4	{"DetectTime": "2021-12-04 ...	2021-12-04T00:00:42Z	df9e654c-4f4a-43b5-92b3-80f370b31007
5	{"DetectTime": "2021-12-04 ...	2021-12-04T00:00:49Z	d7794ed1-fe87-4e14-8718-dd0492f038bb
6	{"DetectTime": "2021-12-04 ...	2021-12-04T00:00:51Z	59f2d3fd-26d8-4fea-a602-43b51a0dd5b1
7	{"DetectTime": "2021-12-04 ...	2021-12-04T00:00:53Z	7f94c19b-0380-4a70-9bba-c12b4a7d33d3
8	{"DetectTime": "2021-12-04 ...	2021-12-04T00:00:55Z	b5d4143d-a3a2-4230-9576-9a72f03d5f5b
9	{"DetectTime": "2021-12-04 ...	2021-12-04T00:00:58Z	6808fea9-0024-417a-a3cd-cf5dbba37900

Obrázok 5: Ukážka súboru z dátovej sady

Každý databázový súbor reprezentuje jeden deň detekcie bezpečnostných udalostí. Je dôležite podotknúť, že časová pečiatka detekcie vyskytujúca sa v našej dátovej sade môže vyskytovať vo forme z časovej zóny CET (Central European Time), pripadne je už upravená do časovej zóny UTC (Coordinated Universal Time). V jednotlivých databázových súboroch sa nachádzajú atribúty ako „id“ pre identifikátor bezpečnostnej udalosti, „time“ ako časová pečiatka a „data“, kde sa nachádza celá bezpečnostná udalosť v IDEA formáte.

3 Geolokačné a reputačné služby

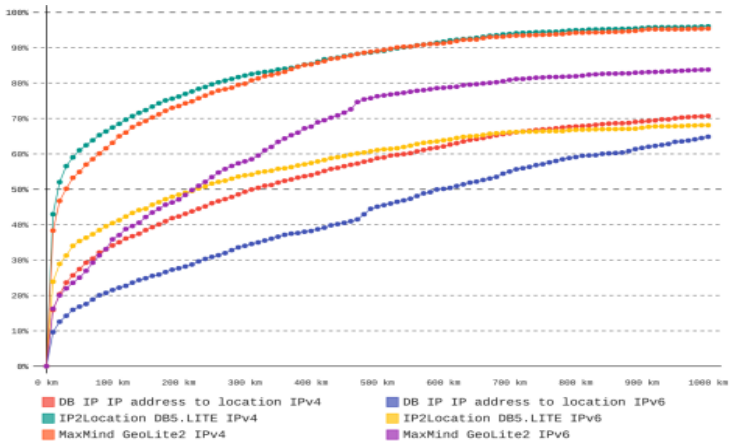
V našej práci sa primárne zameriavame na analýzu z pohľadu zdrojového zariadenia útoku. Z tohto dôvodu je podstatné zistiť približnú lokáciu daného zdrojového zariadenia. Prípadne sa ešte vieme pozrieť na reputáciu zdrojového zariadenia. Tieto faktory nám môžu pomôcť v lepšom pochopení súvislosti pri analýze bezpečnostných udalostí z našej dátovej sady.

3.1 Geolokačné služby

Na to, aby sme mohli začať vysvetľovať geolokačné služby, musíme najprv pochopiť pojem geolokácia. **Geolokácia** v kontexte internetovej komunikácie je problém určovania fyzického umiestnenia alebo lokality (do určitej presnosti) nejakého internetového užívateľa alebo zariadenia [16]. Existuje množstvo geolokačných služieb, ktoré poskytujú geolokačné údaje z IP adresy. Medzi najznámejšie geolokačné databázy patria IP2Location [17], IP-API [18] a DB-IP [19].

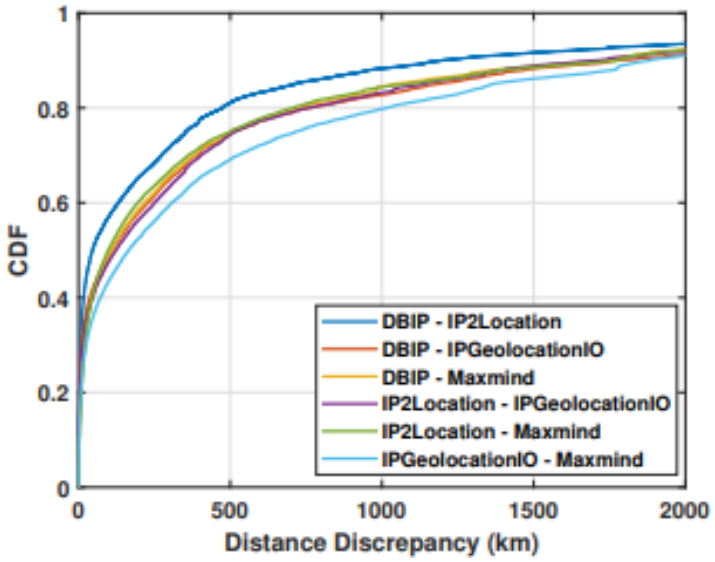
3.1.1 Presnosť geolokačných databáz

Je dôležité podotknúť, že rozdiely medzi jednotlivými službami sú v niektorých prípadoch veľké, či už z hľadiska poskytnutých údajov, ako sú napríklad časová zóna, krajina alebo napríklad poskytovateľ pripojenia k internetu



Obrázok 6: Presnosť geolokačných databáz v rámci daného rozsahu [20]

Na Obr. č. 6 môžeme vidieť presnosť vybraných geolokačných databáz vyjadrených v percentách pre daný rozsah. Z predmetného grafu je možné vidieť, že dané služby vedia poskytnúť len približnú geolokáciu daného zariadenia. Napríklad geolokácia zariadení do 100 km neprekračuje 70% úspešnosť.



Obrázok 7: Presnosť vybraných geolokačných databáz v rámci daného rozsahu [21]

Podobné výsledky sú prezentované aj na Obr. č. 7, kde môžeme vidieť podobnú úspešnosť určenia geolokácie IP adresy pomocou vybraných geolokačných služieb. Je doležíte podotknúť, že pre účely nášho výskumu sú tieto presnosti akceptovateľné, keďže nás primárne zaujíma časová zóna, v ktorej sa útočníkom zneužité zariadenie nachádza

3.1.2 IP-API

My sme sa rozhodli pre využitie IP-API [18] služby pre obohatenie dát z Warden systému o geolokačné informácie. Primárne vďaka jej relatívne vysokej presnosti spolu s tým, že v prípade nedostupnosti záznamu pre jednotlivé IP adresy je z strany IP-API robený dopyt na externú databázu MaxMind [22]. Táto služba poskytuje relatívne presné geolokačné dáta.


```
[{'as': 'AS15169 Google LLC',  'city': 'Ashburn',  'continent': 'North America',  'continentCode': 'NA',
  'country': 'United States','countryCode': 'US',  'countryCode3': 'USA',  'district': '',
  'hosting': True,  'isp': 'Google LLC',  'lat': 39.03,  'lon': -77.5,'mobile': False,
  'org': 'Google Public DNS',  'proxy': False,  'query': '8.8.8.8',  'region': 'VA',
  'regionName': 'Virginia',  'status': 'success','timezone': 'America/New_York',  'zip': '20149'}]
```

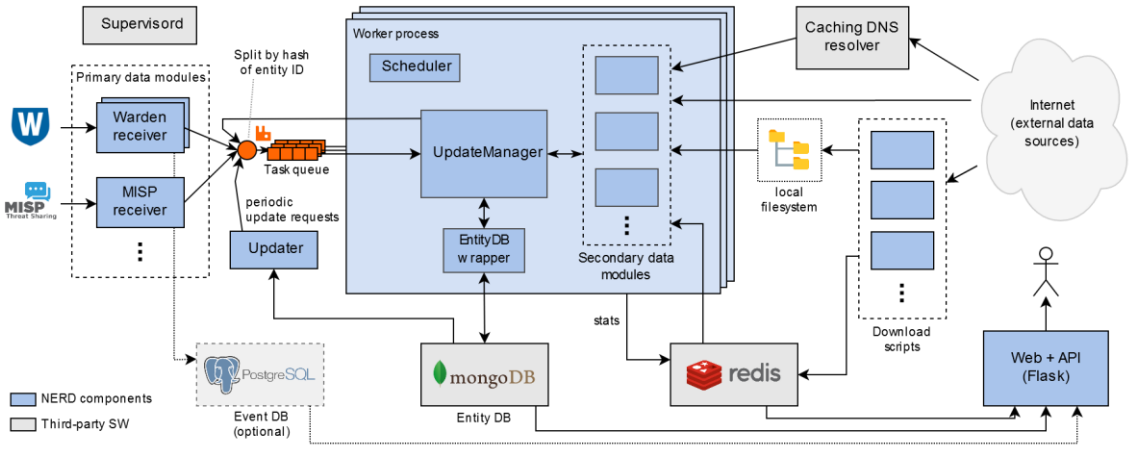

Obrázok 8: Ukážka odpovede IP-API pre IP adresu 8.8.8.8

Rozhodli sme sa implementovať dávkové zapracovanie BATCH JSON. Daná funkcionalita IP-API nám umožňuje vykonať jednu požiadavku pre maximálne 100 IP adries, tak aby sme získali geolokačné údaje pre dané IP adresy. Tento postup sme si vybrali kvôli zrýchleniu procesu napĺňania lokálnej databázy namiesto toho, aby sme požadovali každú IP adresu samostatne.

3.2 Reputačné služby

Pod pojmom **reputácia** si môžeme predstaviť nejaké ohodnotenie zariadenia/entity, podľa nejakých kritérií, ktoré opisujú, akú má dané zariadenie tendenciu vyskytovať sa ako zdroj útoku [23]. Toto ohodnotenie môže byť reprezentované buď číselne, vypočítané podľa určitého vzorca, alebo napríklad len v podobe výskytu na zoznamoch zvaných blacklist. **Blacklist** je zoznam IP adries, ktoré spadajú do kategórie zariadení s vysokou tendenciou vykonávania nejakého špecifického útoku/útokov, pripadne nejakých škodlivých (malicióznych) aktivít [24]. **Reputačné databázy** sú zdroje informácii o škodlivo sa správajúcich entitách z pohľadu informačnej bezpečnosti [23]. **Reputačné skóre** je sumarizácia všetkých známych informácii do jedného čísla reprezentujúceho mieru hrozby, ktorú zariadenia/entita predstavuje [23].

Za externú službu poskytujúcu zoznam výskytov v blacklistoch a reputačné ohodnotenie pre IP adresy z našej dátovej sady sme sa rozhodli použiť službu NERD. Táto služba je, tak isto ako Warden, prevádzkovaná spoločnosťou CESNET. Použitie NERD je pre naše účely výhodné v tom, že výpočet reputácie pre jednotlivé IP adresy sa vykonáva nad bezpečnostnými udalosťami zo služby Warden.

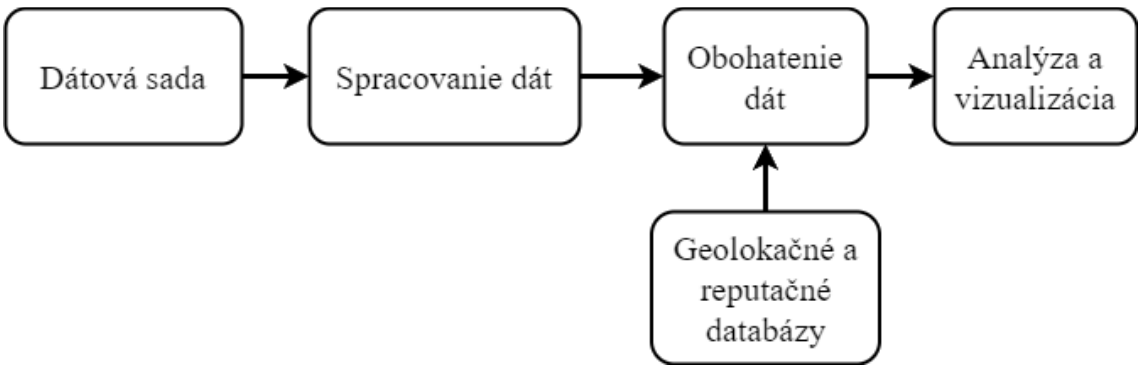


Obrázok 9: Architektúra služby NERD prevzaté z [25].

Na Obr. č. 9 môžeme vidieť náčrt architektúry NERD, ktorá spracováva údaje aj zo systémov ako Warden. Následne dáne údaje upravuje a obohacuje o údaje ako napríklad reputačné skóre a výskyt v blacklistoch pomocou externých služieb. Výpočet reputačného skóre je počítaný na základe denného reputačného skóre a následne upravené ako vážený priemer zo 14 denných skóre. Denné skoré sa počíta pravé podľa výskytu daných IP adries v bezpečnostných udalostiach systému Warden spolu s počtom detekcií na rôznych detekčných zariadeniach [26]. Ako už bolo spomínane NERD taktiež poskytuje zoznam blacklistov, v ktorých sa daná IP adresa nachádza. Výskyt IP adries v blacklistoch NERD sa získava z rôznych externých zdrojov ako napríklad DShield, Spamhaus, AbuseIPDB, atď. [25].

4 Implementácia

Na korektné vykonanie analýzy nad našou dátovou sadou potrebujeme ku každej bezpečnostnej udalosti vedieť čas, kedy ho nejaký systém detegoval a konkrétny indikátor lokality pre útočníkom zneužité zariadenie. Keďže indikátor lokality sa v našich udalostiach nevyskytuje musíme sa uspokojiť s použitím IP adresy, z ktorej neskôr budeme vedieť získať približnú lokalitu zariadenia, ktoré útočník využil na útok. Nakoniec si z jednotlivých udalostí necháme aj použitý protokol komunikácie, číslo portu, na ktorom prebehla komunikácia, a nejaké všeobecné popisy ohľadom daných bezpečnostných udalostí, ako napríklad názov, popis daného pravidla a zariadenia, na ktorom bola daná bezpečnostná udalosť detegovaná.



Obrázok 10: Schéma implementácie riešenia

Na Obr. č. 10 vidíme zjednodušený diagram postupu spracovania, obohatenia a úpravy dátovej sady do podoby v, ktorej je možné aplikovať vybrané prístupy analýz nad danou dátovou sadou. Našu dátovú sadu si musíme najprv spracovať do vhodnej formy na vykonanie analýzy. A teda musíme z databázových súborov bezpečnostných udalosti spraviť súbory v .csv súborovom formáte. Následne z daných súborov vytiahneme IP adresy potrebné na naplnenie lokálnej databázy geolokačnými údajmi. Ak máme lokálnu databázu naplnenú, tak môžeme obohatiť našu dátovú sadu o geolokačné údaje. V poslednom rade môžeme pristúpiť k riešeniu vizualizácie analýzy nad našou dátovou sadou. Pre implementáciu nášho návrhu sme sa rozhodli použiť programovací jazyk Python vo verzii 3.12.1 [27].

4.1 Python a použité nástroje

Programovací jazyk Python sa za posledný pár rokov považuje ako relatívny štandard v dátovej analýze. Medzi hlavné prednosti programovacieho jazyka Python patrí jednoduchá syntax, ktorá umožňuje rýchle učenie a zefektívnenie práce. Python taktiež disponuje veľkým počtom jednotlivých knižníc a modulov, ktoré umožňuje jednoduché spracovanie a analýzu dát. Tieto faktory spolu s veľkou a rastúcou komunitou, robí Python jasnú voľbu pre prácu v oblasti dátovej analýzy [28].

Python disponuje takzvanou Python Standard Library, ktorá je distribuovaná pomocou základnej distribúcie tohto programovacieho jazyka. Táto knižnica ponuka široký sortiment vstavaných modulov, ktoré uľahčujú úlohy ako spracovanie rôznych súborových formátov (napr. csv, json), internetovú komunikáciu (napr. urllib), interakcie s operačným systémom (napr. os, shutil) a oveľa viac. Ak používateľ v tejto knižnici nenájde modul, ktorý by mu vyhovoval, môže si doinštalovať mnoho iných modulov, ktoré sú vytvorené komunitou.

V našej práci sme použili viacero rôznych modulov pre spracovanie, upravenie a vizualizáciu našej dátovej sady, a to:

a) sqlite3 – Sqlite3 je modul, ktorý poskytuje SQL interface pre prácu s SQL databázovými súbormi. Tento modul umožňuje prístup k databázovým súborom s použitím SQL dopytovacieho jazyka bez nutnosti použitia separátneho server procesu [29];

b) ipadress – Tento modul ma schopnosti vytvárať, manipulovať a pracovať s IPv4 a IPv6 adresami a sieťami. Funkcie v tomto module umožňujú užívateľom jednoduchú prácu s IP adresami, overenie platnosti IP adries, overenie či dve IP adresy sú v tej istej podsieti a mnoho ďalších [30];

c) datetime – Datetime modul triedy pre manipuláciu s dátumami a časmi. Taktiež tento modul poskytuje dátumovú a časovú aritmetiku [31];

d) pytz – Modul, ktorý slúži primárne pre výpočty a úpravu dát do časových zón. Rieši problémy s výpočtom dátumu a času pre prechod na a z letného času [32];

e) pandas – Modul Pandas umožňuje užívateľom pracovať s dátami pomocou formátov „Series" a „Dataframe". "Series" je jednorozmerne označené

pole obsahujúce údaje ľubovoľného typu. Dataframe je dvojrozmerná dátová štruktúra, ktorá uchováva údaje ako dvojrozmerné pole alebo tabuľka s riadkami a stĺpcami. Module pandas dokáže priamo pracovať z modulmi datetime a pytz pre spracovanie a prácu s časovými dátami [33];

f) requests – Requests umožňuje jednoduché rozhranie pre prácu s HTTP/1.1 požiadavkami. To zahŕňa odosielanie požiadaviek, manipuláciu s dátami a spracovanie odpovedi [34];

g) json – Modul json v jazyku Python umožňuje rýchlu a jednoduchú prácu s formátom JSON. To zahŕňa rýchlu konverziu dát medzi formátom JSON a Pythonovými dátovými štruktúrami [35];

h) matplotlib –Matplotlib je modul umožňujúci vytváranie statických, animovaných a interaktívnych vizualizácii v Pythone [36];

i) seaborn – Seaborn je modul umožňuje vizualizácie na základe matplotlib. Poskytuje vysoko-úrovňové rozhranie na vytváranie informatívnych grafik [37].

Finálne sme sa rozhodli programovať naše python kódy v prostredí Jupyter Notebook [38]. Toto prostredie nám umožňuje spúšťať jednotlivé časti kódov separátne od celku. Táto funkcia Jupyter Notebooku je výhodná v práci s dátami, lepšej kontroly čiastkových výsledkov a kontroly správnosti uprav dát pre naše spracovanie, obohatenie a analýzu nad našou dátovou sadou.

4.2 Spracovanie dátovej sady

Pomocou nášho kódu sme získali jednotlivé bezpečnostné udalosti z databázových súborov .db, následne sme z IDEA formátu danej bezpečnostnej udalosti vytiahli len nutné parametre, ako je IP adresa, časová pečiatka detekcie, kategória, do ktorej daná udalosť spadá a ďalšie. Keďže naše .db súbory sa nachádzajú v jednom priečinku, získame zoznam týchto súborov, cez ktorý budeme následne iterovať a postupne spracovávať.

```
all_files = glob.glob(os.path.join(path, "*.db"))
```

Následne musíme pristúpiť a čítať z každého .db súboru atribút „data“ z tabuľky „events“. Na tento problém vieme použiť sqlite3 modul, pomocou ktorého získame prístupu k databázovému súboru. Potom využijeme prostredie pandas na spracovanie SQL dopytu v podobe získania údajov z atribútu „data“ tabuľky events.

```
conn = sqlite3.connect(file,isolation_level=None,detect_types=sqlite3.PARSE_COLNAMES)
db_df = pd.read_sql_query("SELECT data FROM events", conn)
```

Ako bolo spomínane jednotlivé údaje v „data“ sú v IDEA formáte, a teda vieme ich následne spracovávať ako JSON objekty. Pandas nám výsledok SQL dopytu uložil ako dátový rámec (dataframe) so stĺpcom „data“. Na to, aby sme upravili daný objekt do dátového rámca s nami zvolenými atribútmi, použijeme nami vytvorenú funkciu „json_data_to_csv(row)“ na dataframe z IDEA udalosťami.

```
db_df["source"], db_df["category"], db_df["detectTime"], db_df["count"], db_df["note"], db_df["description"], db_df["node"], db_df["proto"], db_df["port"] = zip(*db_df.apply(lambda row: json_data_to_csv(row), axis=1))
```

Funkcia „json_data_to_csv(row)“ najprv spracuje záznam v IDEA formáte z atribútu „data“ ako JSON objekt. Pomocou modulu „json“

```
json_data = json.loads(row['data'])
```

Potom naša funkcia nájde jednotlivé údaje v JSON objekte a urobí jednoduchú úpravu a očistenie daného údaja a na konci daný udaj vráti ako reťazec. Ukážeme si to na príklade získania IP adresy zdrojového zariadenia útoku.

```
try:
source=json_data["Source"][0]["IP4"][0].split('/')[0].split('-')[0]
```

```
except Exception as e:
        source=""
```

Vo výsledku sme teda spracovali 40 parametrov z udalostí v IDEA formáte. Následne sme ich celkový počet zredukovali na 9 parametrov. Je dôležité podotknúť, že nutné parametre pre našu časopriestorovú analýzu, sú čas detekcie, časový parameter a IP adresa zariadenia použitého útočníkom, priestorový parameter.

IP	Category	DetectTime	Note	Description	Node	Protocol	Port
5.189.136.235	Attempt.Login+Test	2021-09-20 00:00:00+02:00	Banned by	SSH dictionary/	Fail2Ban	tcp+ssh	22
150.139.212.26	Attempt.Login+Test	2021-09-20 00:00:00+02:00	Banned by	SSH dictionary/	Fail2Ban	tcp+ssh	22
119.28.78.126	Attempt.Login+Test	2021-09-20 00:00:00+02:00	Banned by	SSH dictionary/	Fail2Ban	tcp+ssh	22
211.53.151.151	Attempt.Login+Test	2021-09-20 00:00:00+02:00	Banned by	SSH dictionary/	Fail2Ban	tcp+ssh	22
118.24.120.41	Attempt.Login+Test	2021-09-20 00:00:00+02:00	Banned by	SSH dictionary/	Fail2Ban	tcp+ssh	22
175.27.128.13	Attempt.Login+Test	2021-09-20 00:00:02+02:00	Banned by	SSH dictionary/	Fail2Ban	tcp+ssh	22
58.56.32.238	Attempt.Login+Test	2021-09-20 00:00:02+02:00	Banned by	SSH dictionary/	Fail2Ban	tcp+ssh	22
159.65.160.105	Attempt.Login+Test	2021-09-20 00:00:02+02:00	Banned by	SSH dictionary/	Fail2Ban	tcp+ssh	22

Obrázok 11: Ukážka dátovej sady v súborovom formáte .csv

Na Obr. č. 11 môžeme vidieť malú časť upravených dát z našej dátovej sady, ktorá je pripravená na obohatenie o priestorový udaj a čas zneužitého zdrojového zariadenia.

4.3 Filtrovanie nepoužiteľných bezpečnostných udalostí

Ako už bolo spomínane pri predstavení dátovej sady, naša dátová sada je zlozená z 434 milióna bezpečnostných udalosti. Podstatná časť z tejto dátovej sady sú testovacie udalosti, ktoré boli umelo vytvorené. Z tohto dôvodu musíme z našej dátovej sady odstrániť spomínane testovacie udalosti, udalosti bez zdrojovej IP adresy a bez času detekcie. Na odstránenie udalosti kde sa čas detekcie a zdrojová IP adresa nevyskytujú uplatnime jednoduché príkazy nad našim pandas dataframe-om.

```
db_df = db_df[db_df['source'] != ""]
db_df = db_df[db_df['detectTime'] != ""]
```

Taktiež sme museli z našej dátovej sady odstrániť udalosti, ktorých IP adresa je buď z rozsahu rezervovaných IP adries (napr. 240.0.0.0/4), privátnych IP adries (napr. 192.168.0.0/16, 10.0.0.0/8, ...) alebo multicast IP adries (224.0.0.0/4). Bezpečnostné

udalosti týchto typov IP adries sú pre nás nepoužiteľné, keďže pre IP adresy týchto typov nevieme určiť geolokáciu a ani reputáciu. Na tento problém sme si vytvorili vlastnú funkciu „check_ip_adress(row)“

```
def check_ip_address(row):
return ((ipaddress.ip_address(row['source']).is_global == True and ipaddress.ip_address(row['source']).is_multicast == False))
```

Tuto funkciu aplikujeme na náš dátový rámec a následne odstránime bezpečnostné udalosti, ktoré nespĺňajú nami stanovené podmienky.

```
db_df["private"] = db_df.apply(check_ip_address, axis=1)
db_df = db_df[db_df['private'] != False]
```

Nakoniec nám už stačí odstrániť testovacie bezpečnostné udalosti. Aj na tento problém môžeme použiť funkcionality prostredia pandas v podobe zachovania riadkov, ktoré spĺňajú určitú podmienku.

```
db_df = db_df[db_df['Category'].str.contains('Test') == False]
```

Výsledne po odstránení testovacích a nepoužiteľných udalosti dostávame bezpečnostné udalosti z jedného dňa, ktoré nepatria do skupiny z nežiadúcimi IP adresami alebo bezpečnostné udalosti, ktoré už nie sú umelo vytvorené na účely testovania detekčných zariadení. Celkovo sa teda v našej filtrovanej dátovej sade nachádza 15 rôznych kategórii bezpečnostných udalostí.

Tabuľka 3: Prvých 5 najpočetnejších kategórií bezpečnostných udalostí

Kategória	Počet
Recon.Scanning	183 431 642

Attempt.Login	5 320 568
Intrusion.UserCompromise	5 172 963
Anomaly.Traffic	717 346
Vulnerable.Config	666 699

Z Tab, č. 2 je vidieť, že najpočetnejšia kategória Recon.Scanning tvorí až 94 % všetkých bezpečnostných udalostí. Medzi druhé najpočetnejšie sa radi Attempt.Login s počtom 5,3 milióna bezpečnostných udalostí, tvoriac 3% všetkých bezpečnostných udalostí.

Výsledné atribúty zo spracovaného dátového rámca si uložíme do súboru vo formáte .csv.

```
db_df_out = db_df[['source', 'category', 'detectTime', 'count', 'note', 'description', 'node', 'proto', 'port']]
db_df_out.to_csv(csv_file, index=False, header=False)
```

S takto pripravenou dátovou sadou môžeme začať s obohatením našich dát o približnú lokáciu zdrojového zariadenia, prípadne reputáciu z externých zdrojov ako napríklad IP-API alebo NERD.

4.4 Úprava formátu detekčného času

Ako ďalšiu časť spracovania upravíme detekčný čas do formátu pre časovú zónu UTC. V našej dátovej sade sa detekčný čas nachádza vo viacerých formátoch reprezentujúci čas v určitej zóne. Časový posun pre tuto zónu vieme zistiť z časti, ktorá sa nachádza za "+" v podobe posunu o hodiny a minúty. Napr. „+02:00" je posun o dve hodiny oproti UTC. Niektoré formáty vyjadrujú aj zlomok sekúnd v podobe napr. "2021-09-20 00:01:00.4586+02:00", a teda časť ".4586" reprezentuje zlomok sekundy. Pre naše účely môžeme tieto zlomky vynechať. Na tento účel sme si urobili vlastnú funkciu parse_to_readable_timestamp(timestamp_str), ktorá vyzerá takto:

```
def parse_to_readable_timestamp(timestamp_str):
    parts = timestamp_str.split('+')
    start = parts[0].split(".")
    if len(parts) > 1:
        work_timestamp = start[0]+"+"+parts[1]
    else:
        work_timestamp = start[0]
    if timestamp_str[-1::] == "Z":
        work_timestamp += "Z"
    parsed_timestamp = work_timestamp.replace("ZZ", "Z")
    return parsed_timestamp
```

Odstránime nepotrebné časti (zlomok sekúnd, atď.) a upravíme náš detekčný čas do podoby v časovej zóne UTC. Na dosiahnutie tejto výslednej časovej pečiatky, ktorá bude v štandarde ISO 8601. Použijeme už zabudovanú funkcionalitu prostredia pandas. Túto funkcionalitu sme nepoužili na začiatku, keďže sa v našej dátovej sade vyskytli prípady, ktoré bolo treba upraviť do spracovateľnej formy.

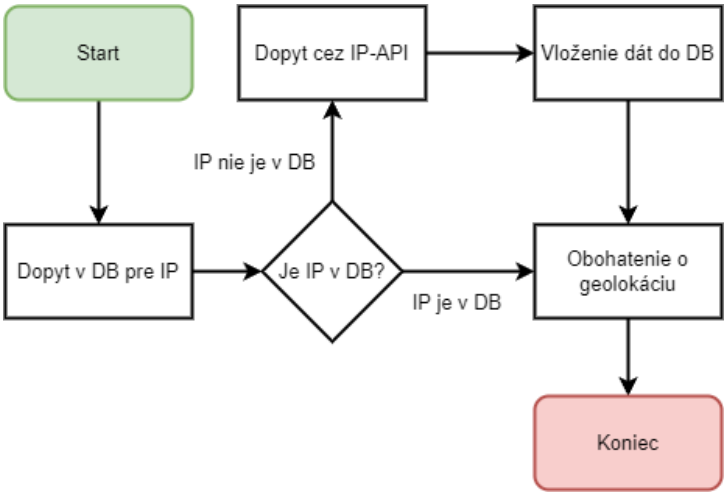
```
df['DetectTime'] = pd.to_datetime(df['DetectTime'], format='mixed', utc=True)
```

Výsledná časová pečiatka bude teda v textovej podobe vyzerať takto „YYYY-MM-DDThh:mm:ssZ“. Táto časová pečiatka je vo formáte podľa normy ISO 8601. Časová pečiatka začína rokom (YYYY), po ktorom nasleduje mesiac (MM) a následne deň (DD) oddelený „-“. Ako ďalší znak nasleduje „T“, ktoré indikuje prechod z dátumovej časti na časovú časť. Časová časť je reprezentovaná hodinou (hh), minútou (mm) a sekundou (ss) oddelenými pomocou „:“. Nakoniec pribudne znak „Z“, ktorý indikuje „Zulu“, a teda časové pásmo UTC.

4.5 Obohatenie o lokalizáciu zariadenia

Na vykonanie časopriestorovej analýzy potrebujeme mať v našej dátovej sade údaj ohľadne času a priestoru. Ako už bolo vyššie spomínané, naša dátová sada disponuje časom detekcie danej bezpečnostnej udalosti, avšak priestorový prvok sa v nej priamo nevyskytuje. Tento údaj vieme získať pomocou IP adresy, a preto musíme navrhnúť postup obohatenia našej dátovej sady pomocou nejakej externej databázy lokality IP adries a naplniť danými údajmi našu lokálnu databázu.

Pre potreby našej práce sme zvolili už spomínanú službu IP-API. Zvolili sme prístup vytvorenia lokálnej databázy IP adries, kde pre každú novú IP adresu z našej dátovej sady budeme uchovávať jej lokalizáciu, a teda pre každú IP budeme musieť žiadať od IP-API jej lokalizáciu.



Obrázok 12: Schéma obohatenia udalosti o lokalizáciu pomocou IP-API

Na Obr. č 12 je načrtnutá schéma obohatenia udalostí o lokalizáciu pomocou vyžiadania lokalizácie pre IP adresu priradenú k danej bezpečnostnej udalosti. Vo výsledku sme sa rozhodli pridať pre každú udalosť nasledujúce údaje z IP-API služby: časová zóna, kde sa daná IP nachádza, krajina výskytu a zemepisná šírka, dĺžka, kontinent, kód kontinentu, kód krajiny, región, názov regiónu, mesto, oblasť, PSČ, poskytovateľ internetového pripojenia, identifikátor organizácie, mobile, proxy a hosting.

Pre lokálnu databázu sme sa rozhodli z dôvodu veľkého počtu rôznych IP adries v našej dátovej sade. Budeme ju využívať na uchovanie geolokačných dát k danej IP adrese. Z týchto dôvodov sme pre naše účely zvolili SQLite_3 modul [29] pre prístup k našej databáze. Pozitívum SQLite_3 modulu je jeho schopnosť pristupovať k databáze

bez vytvorenia server procesu, čo nám spolu s integráciou z jazykom Python, zabezpečí ľahký prístup k dátam uložených v našej databáze. Najprv si teda vytvoríme danú databázu spolu z indexom pre lepšie vyhľadávanie v databáze.

```
con = sqlite3.connect(path)
cur = con.cursor()
cur.execute("CREATE TABLE ipapitable(query,continent,continentCode,country,countryCode,countryCode3,region,regionName,city,district,zipCode,lat,lon,timezone,isp,org,asnum,mobile,proxy,hosting)")
cur.execute("CREATE INDEX idx_query ON ipapitable(query)")
con.commit()
```

Našu lokálnu databázu naplníme geologickými dátami podľa IP adries nachádzajúcich sa v našej dátovej sade pomocou už spomínanej externej služby IP-API.

```
for chunk in pd.read_csv(file_path, chunksize=100):
    try:
        ip_list = chunk.iloc[:, 0].tolist()
        json_array = json.dumps(ip_list)
 headers = {'Content-Type': 'application/x-www-form-urlencoded',}
        params = {'fields': '87781375','key': 'KEY',}
        response = requests.post('https://pro.ip-api.com/batch', params=params, headers=headers, data=json_array)
```

Následné si z danej odpovede získame jednotlivé atribúty , ktoré sme si vyžiadali a pridáme ich do lokálnej databázy pomocou príkazu „cur.executemany()”.

```
cur.executemany("INSERT INTO ipapitable VALUES(?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?)", data_out)
 con.commit()
```

Celkovo sa v našej práci nachádza 5,37 milióna jedinečných IP adries. My sme sa rozhodli rozdeliť našu databázu na 12 celkov. Každý celok reprezentuje jeden mesiac údajov. To sme vykonali najmä z dôvodu veľkosti databázy a urýchlenia vyhľadávania v databáze.

query	continent	country	lat	lon	timezone	mobile	proxy	hosting
Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
202.91.78.34	Asia	India	30.7276	78.4443	Asia/Kolkata	0	0	0
172.86.125.78	North America	United States	34.0726	-118.261	America/Los_Angeles	0	1	0
45.66.137.201	North America	United States	32.7797	-96.8022	America/Chicago	0	1	0
200.142.168.253	South America	Brazil	-22.8974	-43.1803	America/Sao_Paulo	0	0	0
177.249.47.241	North America	Mexico	20.8214	-103.4595	America/Mexico_City	0	0	0

Obrázok 13: Ukážka časti tabuľky z lokálnej databázy pre obohatenie pomocou IP-API

Na Obr. č. 13 môžeme vidieť ukážku časti tabuľky z našej databázy zodpovedajúcej geolokačným dátam pre IP adresy. Celkovo sa v našej lokálnej databáze nachádza 20 rôznych atribútov. Ide o tieto atribúty: „query", „continent", „continentCode", „country", „countryCode", „countryCode3", „region", „regionName", „city", „district", „zipCode", „lat", „lon", „timezone", „isp", „org", „asnum", „mobile", „proxy" a „hosting". Medzi najvýznamnejšie atribúty pre našu analýzu patrí „lat" a „lon" určujúcu približnú zemepisnú šírku a dĺžku, atribút „country" určujúci krajinu výskytu, atribút „timezone" určuje časovú zónu, v ktorej sa zariadenie nachádza a atribúty „mobile", „proxy" a „hosting" opisujúce typ zariadenia, ktoré je na danej IP adrese (či ide o mobilné zariadenie, proxy server alebo server poskytujúci služieb hosting).

4.6 Obohatenie o čas zdrojového zariadenia

Naša dátová sadu disponuje časom detekcie. Ak sa chceme zamerať na čas zdroja útoku (zneužitého zariadenia, ktoré využil útočník na útok), tak potrebujeme najprv vedieť jeho približnú lokáciu. Touto lokáciou už disponujeme, keďže sme si už naplnili našu lokálnu databázu geolokačnými dátami. Následne už vieme pre každú bezpečnostnú udalosť vypočítať jej čas zdrojového zariadenia a pridať danú hodnotu. Pomocou časovej pečiatky detekcie, ktorá sa nachádza v časovej zóne UTC (Coordinated Universal Time) a časovej zóny získanej z našej lokálnej databázy, vieme jednoducho vypočítať čas zdrojového zariadenia. Keďže náš dátový rámec akceptuje detekčný čas ako datetime

objekt, môžeme využiť modul pytz a upraviť daný datetime objekt do datetime objektu reprezentujúceho čas zdrojového zariadenia pomocou atribútu „timezone“.

```
def get_sourceTime_datetime(row):
return
row['DetectTime'].astimezone(row["timezone"]).strftime('%Y-%m-%d
%H:%M:%S')
```

Táto funkcia nám vypočíta čas zdrojového zariadenia pomocou časovej zóny získanej z IP-API [18] a upraví danú časovú pečiatku do podoby reťazca. Nakoniec, tento čas zdrojového zariadenia vypočítaného z detekčného času pridáme do našej dátovej sady.

IP	Category	DetectTime	Description	Node	Protocol	Port	lat	lon	timezone	SourceTime
63.81.91.50	Abusive.Spam	2021-08-31T22:01:44Z	Blacklisted host	Fail2Ban	NaN	NaN	38.5764	-121.3070	America/Los_Angeles	2021-08-31 15:01:44
63.81.84.18	Abusive.Spam	2021-08-31T22:47:36Z	Blacklisted host	Fail2Ban	NaN	NaN	38.5764	-121.3070	America/Los_Angeles	2021-08-31 15:47:36
37.123.101.98	Abusive.Spam	2021-08-31T22:51:45Z	Blacklisted host	Fail2Ban	NaN	NaN	38.6133	27.3724	Europe/Istanbul	2021-09-01 01:51:45
63.81.84.124	Abusive.Spam	2021-08-31T23:08:01Z	Blacklisted host	Fail2Ban	NaN	NaN	38.5764	-121.3070	America/Los_Angeles	2021-08-31 16:08:01
63.80.190.65	Abusive.Spam	2021-08-31T23:30:48Z	Blacklisted host	Fail2Ban	NaN	NaN	38.5764	-121.3070	America/Los_Angeles	2021-08-31 16:30:48
37.123.101.112	Abusive.Spam	2021-08-31T23:51:19Z	Blacklisted host	Fail2Ban	NaN	NaN	38.6133	27.3724	Europe/Istanbul	2021-09-01 02:51:19
194.163.160.114	Recon.Scanning	2021-09-01T00:00:00Z	Horizontal port scan	Nemea	NaN	NaN	51.1878	6.8607	Europe/Berlin	2021-09-01 02:00:00

Obrázok 14: Ukážka upravených dát s pridaným časom zdrojového zariadenia

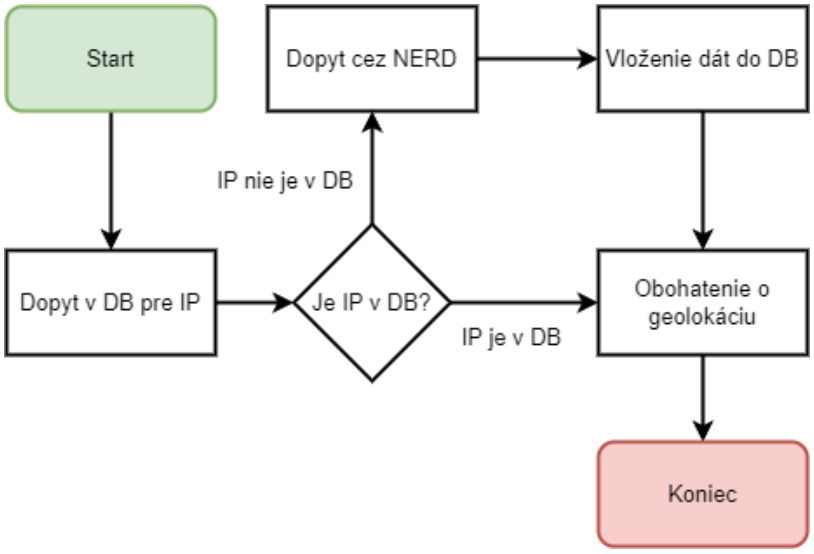
Na obrázku č. 14 môžeme vidieť ukážku časti obohatenej dátovej sady. Spolu s už vypočítaním a pridaným časom zdrojového zariadenia. Pravé tento atribút je podstatný pre našu prácu, keďže sme sa zamerali na analýzu bezpečnostných udalosti z pohľadu zdrojového zariadenia, ktoré bolo zneužite na útok.

4.7 Obohatenie o reputáciu a typ zariadenia

Keďže plánujeme vykonať analýzy nad našou dátovou sadou aj podľa reputácie, typu, či výskytu zdrojového zariadenia v blacklistoch, musíme prvotne získať tieto údaje. Určenie typu zariadenia pre jednotlivé IP adresy sme už získali zo služby IP-API, a teda v prípade typov zariadení nám už len stačí tieto údaje vložiť do našej dátovej sady. V prípade reputácie a výskytu v blacklistoch sme sa rozhodli vykonať analýzu nad dátovou sadou z Apríla 2024. Medzi hlavne dôvody patri, už v predošlých častiach spomínaný,

vzorec pre výpočet reputácie v službe NERD, ktorý počíta reputáciu pre danú IP adresu iba za posledných 14 dni.

Princíp obohatenia ostáva takmer taký istý ako u získania údajov z IP-API. A teda najprv si naplníme lokálnu databázu údajmi zo služby NERD, kde po naplnení vložíme údaje o reputácii a blacklistoch pre jednotlivé IP adresy do našej dátovej sady.



Obrázok 15: Schéma obohatenia bezpečnostnej udalosti o lokalizáciu pomocou NERD

Na Obr. č 15 je načrtnutá schéma obohatenia bezpečnostnej udalosti o reputáciu pomocou vyžiadania lokalizácie pre IP adresu. Vo výsledku sme sa rozhodli pridať pre každú udalosť nasledujúce údaje z NERD: reputačné skóre a zoznam blacklistov, v ktorých sa IP adresa nachádza.

Pre zjednodušenie práce a sprehľadnenie sme sa rozhodli upraviť už nami vytvorenú infraštruktúru. Tak ako aj u databázy s geolokačnými dátami aj u tejto z reputačnými dátami využijeme Sqlite3 modul pre prístup k nej. Podobne ako u naplnení databázy geolokačnými dátami aj u reputačných musíme vytvoriť správny databázový súbor spolu z tabuľkou a indexom.

```
con = sqlite3.connect(path)
cur = con.cursor()
cur.execute("CREATE TABLE nerdIPinfo(query,rep,hostname,black)")
cur.execute("CREATE INDEX idx_query ON nerdIPinfo(query)")
```

```
con.commit()
```

Naša lokálna databáza, pre reputačné skóre a výskyt v blacklistoch, ma len štyri atribúty a to atribút „query“ pre IP adresu, atribút „rep” pre reputačné skóre, atribút „hostname“ pre názov zdrojového zariadenia a atribút„black“ pre zoznam blacklistov, v ktorých daná IP adresa figuruje.

```
for chunk in pd.read_csv(path_IP, chunksize=1, usecols=[0], header=None):
    try:
        ip_list = chunk.iloc[:, 0].tolist()
        data = ','.join(ip_list).strip()
        headers = {
        'Authorization': 'KEY',
        }
     response = requests.get('https://nerd.cesnet.cz/nerd/api/v1/ip/'+str(data.strip()), headers=headers)
        data_out= []
```

Pomocou tohto kódu sme získali údaje ohľadom jednej IP adresy. Následne tieto údaje musíme spracovať a upraviť do použiteľnej formy.

```
query = str(response.json()['ip'])
rep = str(response.json()['rep'] )[:5]
hostname = response.json()['hostname']
black = "+".join(response.json()['bl'])
data_out.append((query,rep,hostname,black))
```

Takto upravene údaje vložíme do našej lokálnej databázy pomocou príkazu „cur.executemany()”.

```
        cur.executemany("INSERT INTO nerdIPinfo VALUES(?, ?, ?, ?)",
data_out)
        con.commit()
```

query	rep	hostname	black
Filter	Filter	Filter	Filter
206.168.34.170	0.955	unused-space.coop.net	ciarmy+abuseipdb+dshield+turris_greylist
206.168.34.160	0.947	unused-space.coop.net	ciarmy+abuseipdb+dshield+turris_greylist
206.168.34.174	0.954	unused-space.coop.net	ciarmy+abuseipdb+dshield+turris_greylist
206.168.34.165	0.949	unused-space.coop.net	ciarmy+abuseipdb+dshield+turris_greylist
206.168.34.167	0.954	unused-space.coop.net	ciarmy+abuseipdb+dshield+turris_greylist
206.168.34.169	0.947	unused-space.coop.net	ciarmy+abuseipdb+dshield+turris_greylist
206.168.34.171	0.947	unused-space.coop.net	ciarmy+abuseipdb+dshield+turris_greylist
206.168.34.163	0.948	unused-space.coop.net	ciarmy+abuseipdb+dshield+turris_greylist
206.168.34.162	0.949	unused-space.coop.net	ciarmy+abuseipdb+dshield+turris_greylist

Obrázok 16: Ukážka časti tabuľky z lokálnej databázy pre obohatenie pomocou NERD

Na Obr. č. 16 môžeme vidieť ukážku tabuľky z našej databázy pre reputačné dáta zo služby NERD. Celkovo sa v našej tabuľke nachádzajú len štyri atribúty, keďže najpodstatnejšie je pre nás reputačné skóre a zoznam výskytu v blacklistoch. Zoznam výskytu blacklistov uchovávame vo forme reťazca s oddeľovačom v podobe „+“. Výsledne máme teda v tabuľke atribúty „query“ pre IP adresu, „rep“ pre reputačné skóre, „hostname“ pre názov zariadenia na IP adrese a „black“ pre zoznam blacklistov.

4.8 Finálne úpravy

Ako posledné sme sa rozhodli pridať par atribútov do našej už spracovanej a obohatenej dátovej sady za účelom zrýchlenia výpočtu analýz. Medzi takéto atribúty patri napríklad „source_day_of_week“, ktorý reprezentuje deň v týždni kedy sa vykonal útok z zneužitého zdrojového zariadenia. Tento udaj sme získali z nami získaného zdrojového času pomocou funkcie strftime, keďže náš dataframe daný atribút už akceptuje ako datetime objekt. Tato funkcia nám umožňuje získať deň v týždni z datetime objektu.

```
def get_day_of_the_week_from_sourceTime(row):
    return row["SourceTime"].strftime('%A')
```

Na uplatnenie funkcie pre nový stĺpec „source_day_of_week“ aplikujeme nasledovný príkaz.

```
df['source_day_of_week'] = df.apply(get_day_of_the_week_from_sourceTime, axis=1)
```

Viacero atribútov ako napríklad „source_hour“ a „source_day“ a ďalšie, sme získali pravé pomocou našich vlastných funkcii ako napr. hour.

```
def hour(timestamp):
    return str(timestamp)[11:13]
```

Dané funkcie, pre jednotlivé atribúty aplikujeme podobne ako u „source_day_of_week“.

```
df['source_hour'] = df['SourceTime'].apply(hour)
```

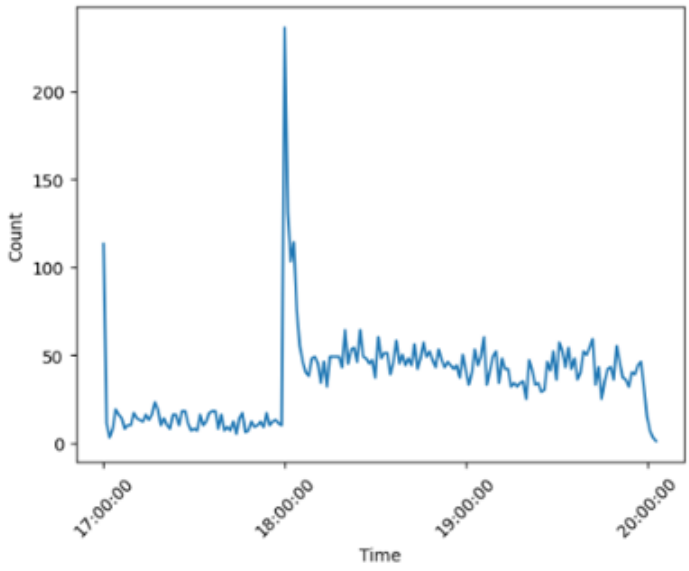
Celkovo máme 12 .csv súborov (11 pre rok 2021 a 1 pre apríl 2024), ktoré disponujú celkovo 28 atribútmi. Okrem už spomínaných 9 atribútov zo spracovania dátovej sady („IP“, „DetectTime“, „Category“, atď.) nám pribudlo ešte 19 atribútov z procesu obohatenia a finálnych úprav („lat”, „lon”, „timezone”„Sourcetime”, „nerd_rep”, atď.). V konečnom dôsledku táto spracovaná, obohatená a upravená dátová sada je priprávaná na analýzu.

5 Analýza a vizualizácia

Na korektné pochopenie výsledkov časopriestorovej analýzy je nevyhnutné korektne implementovať vizualizáciu výsledkov v ľudsky čitateľnej a pochopiteľnej forme. Z týchto dôvodov sme sa rozhodli použiť časove rady a tepelné mapy na vizualizáciu výsledkov analýzy a nájdenie vzorov medzi udalosťami.

5.1 Časové rady

Časový rad je súbor nejakých meraní/pozorovaní x_t, kde pre každé x_t platí, že bolo zaznamenané v nejakom určitom čase t“ [39]. Z tejto definície je jasné, že pre naše účely sa daná forma vizualizácie ľahko implementovateľná, keďže pre naše pozorovania (bezpečnostné udalosti) platí, že každá udalosť má konkrétnu časovú pečiatku ich detekcie. A jednoduchým rozdelením našej dátovej sady na skupiny podľa časových zón získaných z procesu obohatenia dát o priestorový parameter, vieme uplatniť časové rady na každú z časových zón.



Obrázok 17: Príklad časového radu z testovacích dát

Takto dokážeme vizualizovať trend bezpečnostných udalostí podľa časových zón v reprezentácii počtu udalostí podľa času v danej časovej zóne, alebo aj podľa iných časových a priestorových metrík ako je napríklad samotný lokálny čas zariadenia získaný pomocou času detekcia a lokácie zariadenia. Príkladom tohto je aj Obr. č. 17, v ktorom

môžeme vidieť nárast bezpečnostných udalostí v priebehu dňa z vybraného úseku testovacích dát. Konkrétne v tomto prípade daný časový rad poukazuje na anomáliu v počte udalosti, kde okolo 18:00 prudko stúpol počet bezpečnostných udalosti.

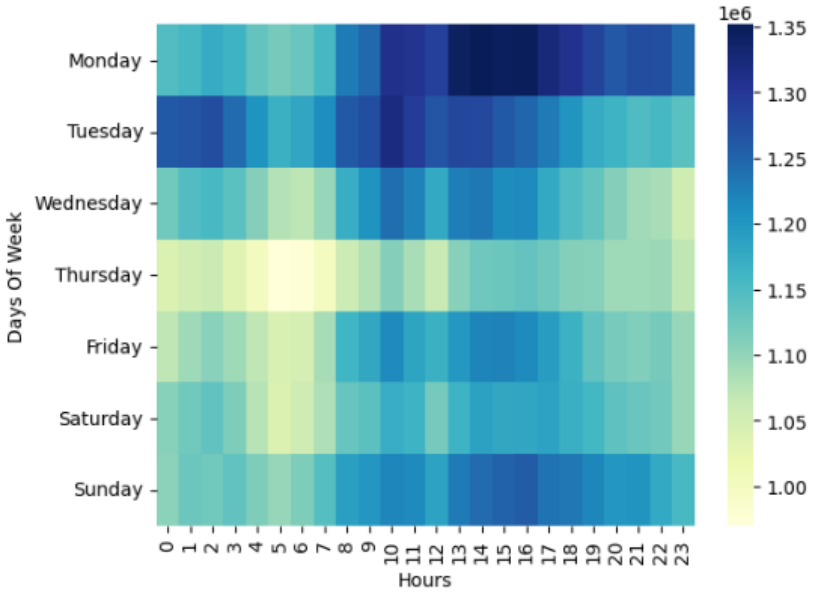
5.2 Tepelné mapy

Tepelné mapy (heatmaps) predstavujú dvojrozmerné tabuľky čísel reprezentovane nejakým farebným odtieňom [40]. Rozhodli sme sa použiť tepelné mapy, pretože nám umožňujú zobrazovať geografické údaje rozumnejším spôsobom. Pridaním časových údajov a následným filtrovaním podľa času môžeme vykonať časovo-priestorovú analýzu and našou dátovou sadou. To nám môže pomôcť lepšie vizualizovať údaje z našej dátovej sady, ktoré boli obohatené o geolokačné a časové údaje.

Keďže naša dátová sada má veľkosť 56 GB po úprave a odstránení testovacích dát , musíme nájsť spôsob ako postupne vypočítavať priebežné výsledky súbor po súbore. Tento postup uplatníme na celú dátovú sadu s využitím modulu pandas a jeho funkcionality pivot tabuliek.

```
      pivot_df4 = pd.pivot_table(df4, index='source_day_in_the_week',
columns='source_hour', aggfunc='size', fill_value=0)
      pivot_combined = pivot_combined.add(pivot_df4, fill_value=0
      sns.heatmap(pivot_combined, cmap="YlGnBu")
      plt.ylabel('Reputaion Quadrant')
      plt.xlabel('SourceHour')
      plt.show()
```

Nakoniec sme na vykreslenie tejto tepelnej mapy použili Python modul seaborn. Tento modul nám umožnil intuitívne meniť vzhľad tepelnej mapy pomocou rôznych farebných paliet a nastavení.



Obrázok 18: Ukážka tepelnej mapy pre celú dátovú sadu

Na príklade z Obr. č. 18 môžeme vidieť tepelnú mapu vytvorenú nad dátovou sadou pre rok 2021. Táto mapa reprezentuje vzťah medzi dňom v týždni, v ktorom sa zdrojové zariadenie nachádzalo počas útoku a hodinou v dni.

5.3 Analýza nad celou dátovou sadou

V predošlých častiach tejto práce sme popísali hlavné postupy a princípy použité pri spracovaní dátovej do vizuálnej formy tepelných a časových radov. V tejto časti sa budeme venovať analýze dátovej sady a následnej vizualizácii našich výsledkov analýzy.

5.3.1 Lokácia zneužitých zariadení

Ako ďalšie sme sa pozreli na rozdelenie bezpečnostných udalosti podľa krajín a časových zón, v ktorých sa zdroj útoku (zneužité zariadenie) nachádza.

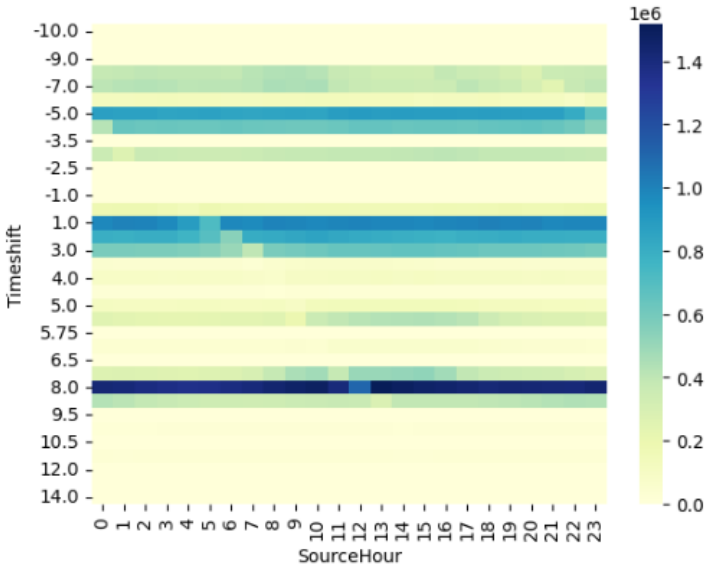
Tabuľka 4: Prvých 5 najpočetnejších krajín z bezpečnostných udalosti

Krajina	Počet
United States	50 295 744

China	23 719 157
Russia	11 718 589
The Netherlands	11 080 055
Brazil	8 206 338

Z Tab. č. 3 je jasne vidieť, že výskyt krajín USA, Čína a Rusko je v našich dátach dominantný. Tento trend vychádza z distribúcie IPv4 rozsahov adries medzi jednotlivé krajiny a kontinenty, kde krajina s najväčším počtom pridelených IP adries je USA.

Teraz sa pozrieme na tepelnú mapu vysvetľujúcu vzťah medzi hodinami a časovým posunom z dôvodu časovej zóny. Vyššie v práci sme uviedli, že predpokladáme, že zdrojom útoku sú zneužité zariadenia. Uvedenie jednotlivých krajín teda nepredstavuje nepriateľské tendencie danej krajiny, ale množstvo zneužitých zariadení (približne) umiestnených v danej krajine.



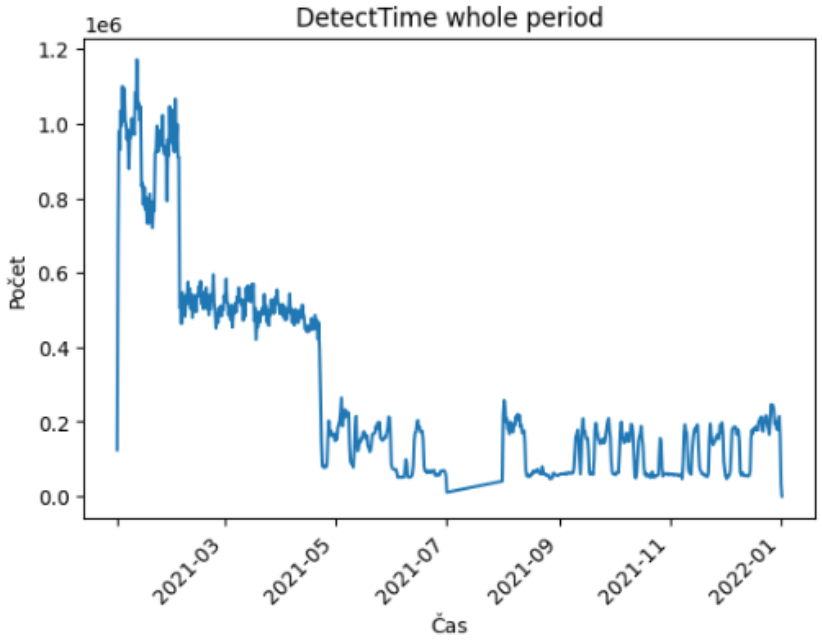
Obrázok 19: Tepelná mapa nad časovým posunom a hodinou.

Dominantnosť krajín z Tab. č. 2 sa potvrdila aj v tepelnej mape z Obr. č. 19. Vidíme, že časový posun +8 zastupuje Čínu, z dôvodu, že Čína používa na celom svojom území fixný časový posun +8. Taktiež môže vidieť silnú koncentráciu pri časovom posune +1 a +3 čo odpovedá krajinám Holandsko a Rusko (oblasť okolo Moskvi).

Spojene štáty majú naprieč svojim územím 4 časové zóny, pričom v našej tepelnej mape to môžeme vidieť ako zvýšení počet udalosti z časovým posunom -7, -8 odpovedajúce západnému pobrežiu spojených štátov a -4, -5, ktoré reprezentujú východne pobrežie Spojených štátov.

5.3.2 Analýza podľa času detekcie útoku na detekčnom zariadení

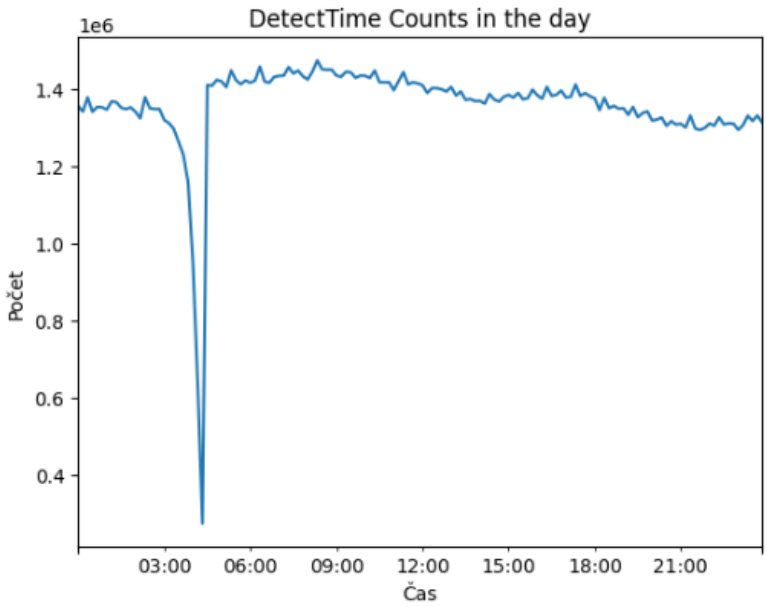
Keďže sme sa v práci zamerali na čas zdroja útoku a taktiež chceme poukázať na významnosť daného spôsobu, musíme najprv urobiť analýzy z pohľadu detekčného času. Tato skutočnosť nám umožní porovnať analýzy z pohľadu obete na ktorú útočník cieli.



Obrázok 20: Časový rad počtu bezpečnostných udalostí z celej dátovej sady (čas detekcie).

Na obrázku č. 20 môžeme vidieť reprezentáciu našej dátovej sady pomocou časového radu za celé obdobie zachytené v našej dátovej sade (rok 2021-2022). Ako časovú pečiatku sme použili čas detekcie cieľom útoku. Z našej dátovej sady je zrejmé, že prvý mesiac január disponoval najväčším počtom bezpečnostných udalostí, keď v jednom úseku 12 hodín dosiahol počet takmer 1.2 milióna udalostí. Postupne vidíme pokles počtu udalostí na stabilnú úroveň v mesiacoch február až apríl 2021. Nakoniec nám počet udalostí klesol až na maximálne 240 tisíc udalostí za 12 hodín v priebehu mesiacov máj až december 2021.

Ako ďalšie sa pozrieme na výsledok analýzy počtu bezpečnostných udalostí podľa času detekcie v dni. Bude nás teda zaujímať ako pribúdajú bezpečnostné udalosti v priebehu dňa.

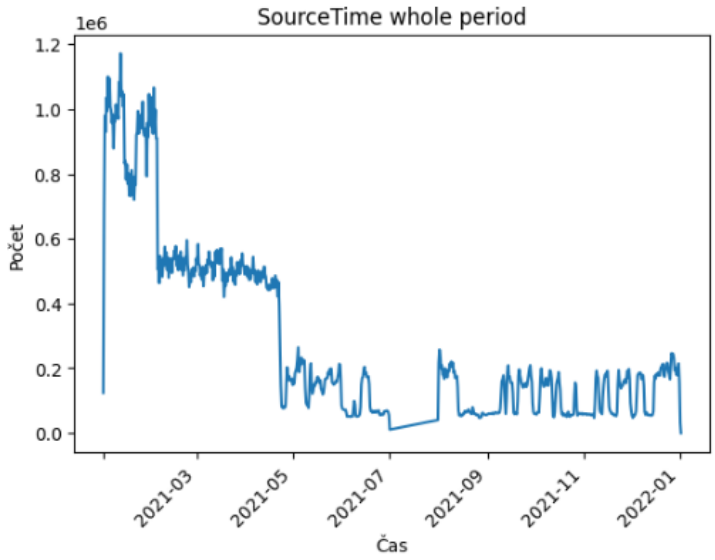


Obrázok 21: Časový rad počtu bezpečnostných udalostí v dni z celej dátovej sady (čas detekcie).

Na obrázku č. 21 vidíme výsledok analýzy, kde sme sa zamerali na počet bezpečnostných udalostí počas dna. Môžeme vidieť, že z pohľadu cieľa útoku bezpečnostné udalosti pribúdali konštante s výnimkou okolo 5. hodiny ráno. V tomto momente vidíme pokles zo štandardných takmer 1,4 milióna udalostí za 10 min na 300 tisíc bezpečnostných udalostí. Po analýze viacerých parametrov sme dospeli k záveru, že dané správanie je výsledkom pravdepodobného problémového správania systému Warden v danom čase. Vo väčšine atribútov, ako napríklad v počte udalostí detegovaných na detekčných zariadeniach (honeypotoch) Cowrie a Nemea, je daný pokles identický.

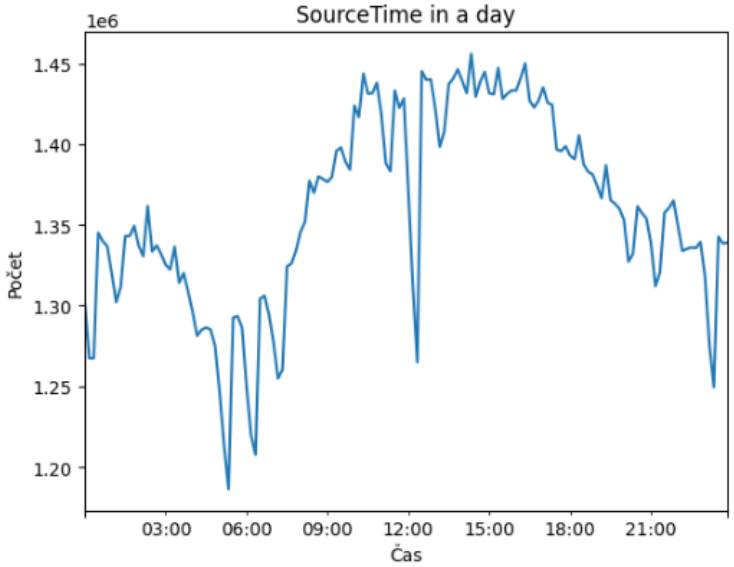
5.3.3 Analýza podľa času útoku zdrojového zariadenia

Ako už bolo spomínane v predošlých častiach, našu analýzu sme zamerali primárne na čas zdroja útoku (zneužitého zariadenia), ktorý sme získali pomocou časovej zóny z externej služby IP-API pre dané zariadenie a času detekcie útoku daným detekčným zariadením.



Obrázok 22: Časový rad počtu bezpečnostných udalostí z celej dátovej sady (čas zdrojového zariadenia)

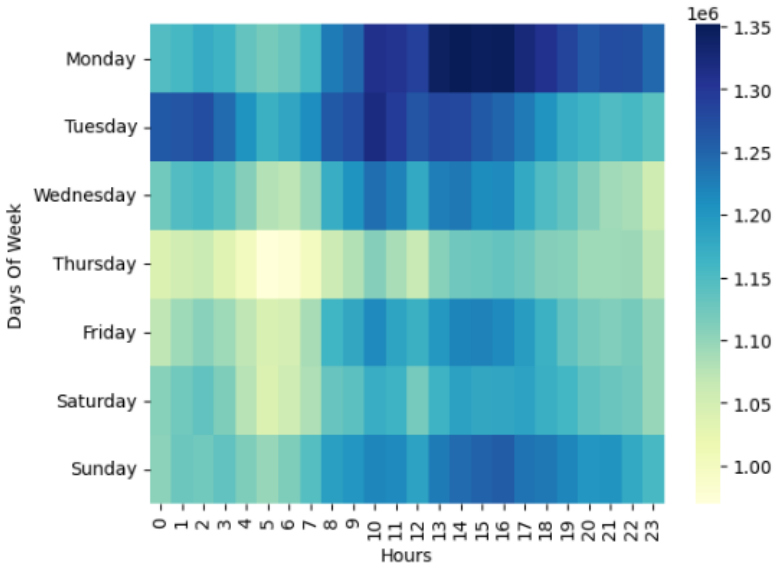
Na obrázku č. 22 vidíme výsledok podobnej analýzy ako na obrázku č. 20, avšak v tomto prípade sme použili nami obohatený atribút, a to čas zdrojového zariadenia. Tento graf je podobný grafu na obrázku č. 20, keďže časový rozdiel medzi detekčným časom a časom na zdrojovom zariadení sa môže líšiť maximálne o 26 hodín, kvôli časovým zónam (UTC-12 a UTC+14) a posunu pre letný alebo zimný čas.



Obrázok 23: Časový rad počtu bezpečnostných udalostí v dni z celej dátovej sady (čas zdrojového zariadenia).

O mnoho zaujímavejšia je analýza, ktorá sa zamerala na počet bezpečnostných udalosti počas dňa. Tento typ analýzy nám môže pomôcť odhaliť špecifický čas, počas ktorého môžu byt zneužité zariadenia viac zneužívane ako zvyčajne. Túto analýzu môžeme vidieť na Obr. č. 23. Oproti analýze podľa času detekcie, ktorú môžeme nájsť na Obr. č. 21, je jasne vidieť diametrálne odlišný trend počtu bezpečnostných udalostí. Môžeme vidieť poklesy využitia zariadení na útok v časoch okolo 6:00 až 8:00, tesne po 12:00 a v nočných hodinách okolo 23:00.

Útočníci na svoje útoky využívajú zariadenia , ktorých majitelia v drvivej väčšine nevedia o zapojení ich zariadenia do nejakého útoku (útočníkom zneužité zariadeni). Dané poklesy môžu nasvedčovať majiteľovej aktivite so svojim zariadením, čo vedie útočníkov k pozastaveniu využívania daného zariadenia na útok a vyhnutiu sa detekcie.



Obrázok 24: Tepelná mapa nad dňom v týždni a hodinou (čas zdrojového zariadenia).

Na obrázku č. 24 môžeme vidieť výsledok analýzy nad dňom v týždni a hodinou zdrojového zariadenia, kedy sa odohral útok na detekčnom zariadení. Môžeme vidieť, že v našej dátovej sade sa útoky z pohľadu zneužitého zariadenia vo všeobecnosti vykonávali v poobedňajších hodinách s poklesmi o 12. a 5. až 6. hodine. Najčastejšie sa útoky vykonávali v pondelok v poobedňajších hodinách, utorok okolo 1. až 2. hodine ráno, utorok okolo 9. až 11. a v nedeľu okolo 13. až 16. hodine poobede. Ak sa zameriame pravé na pondelok a utorok, zistíme že zvýšený počet bezpečnostných udalosti v nasej dátovej sade úzko kopíruje trendy obsiahnuté v celkových analýzach. Percentuálna distribúcia jednotlivých kategórii a krajín zdrojových zariadení je zachovaná a teda

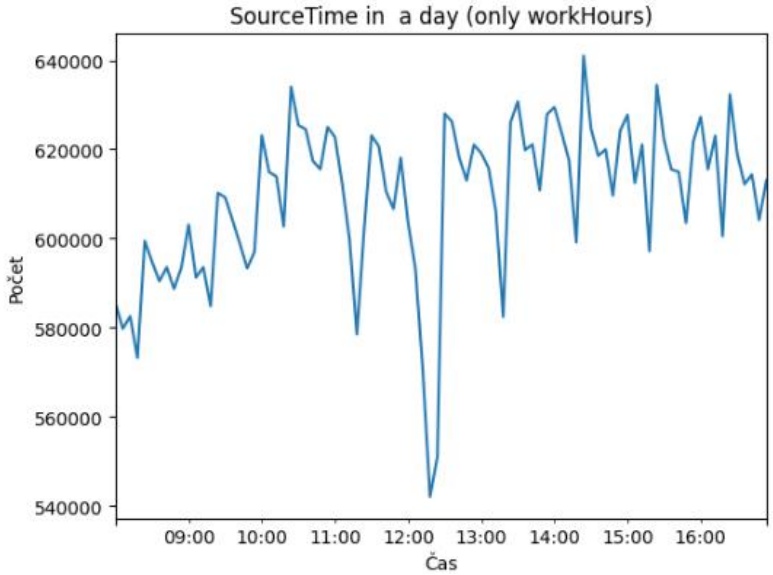
môžeme konštatovať, že zvýšená aktivita útočníkov v daných dňoch súvisy z celkovým trendom útočníkov využívať zneužité zariadenia v daných dňoch.

5.4 Analýza vzhľadom na špecifické podmnožiny

V predchádzajúcom texte sme popísali, ako vyzerá naša dátová sada. Vykonali sme základne analýzy nad celou dátovou sadou a pozreli sme sa na počty jednotlivých atribútov v našej dátovej sade. Pričom sme vizualizovali jednotlivé atribúty do časových radov a tepelných máp zo zameraním na čas zdrojového zariadenia, ktoré bolo zneužite na útok. V rámci tejto podkapitoly sa pozrieme na podobné analýzy, avšak našu dátovú sadu budeme filtrovať pomocou určitých parametrov ako napríklad lokácia, kategória, reputácie skóre a mnoho ďalších, pričom nakoniec porovnáme jednotlivé analýzy oproti výsledkom z celkovej analýzy.

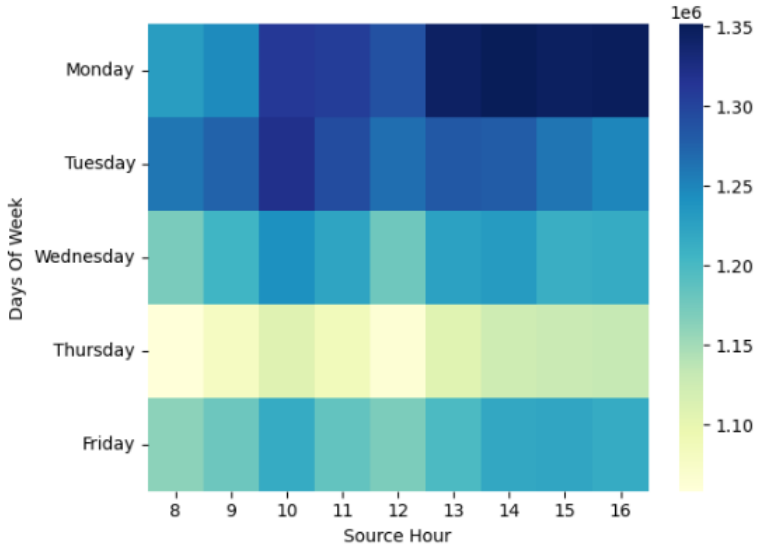
5.4.1 Rozdelenie bezpečnostných udalosti podľa pracovného času

V tejto časti sme sa rozhodli zamerať na pracovnú dobu a počty jednotlivých bezpečnostných udalostí v tejto pracovnej dobe. Pracovnú dobu sme si definovali od 8:00 do 17:00. Pričom sa zameriavame na čas zdrojového zariadenia a vplyv tohto času na počty bezpečnostných udalosti za daný časový úsek.



Obrázok 25: Časový rad počtu udalostí v dni počas pracovnej doby (čas zdrojového zariadenia).

Z analýzy na obrázku č. 25 je zrejmé, že počty bezpečnostných udalostí z pohľadu zdrojového zariadenia v našej dátovej sade klesali v časovom úseku okolo 12:00 až 13:00. Toto správanie môže nasvedčovať zvýšenej aktivite držiteľa zneužitého zariadenia z dôvodu, napr. obedňajšej prestávky. Taktiež môžeme na Obr. č. 25 vidieť systematický rast a pokles bezpečnostných udalostí v jednohodinových intervaloch. Toto správanie môže nasvedčovať automatizácii z pohľadu útočníka, ktorého skript/kód je spúšťaný v jednohodinových intervaloch. To môže spôsobiť snahu útočníkov vyhnúť sa detekcii a následnému zníženiu aktivity na danom zariadení z pohľadu útočníka.

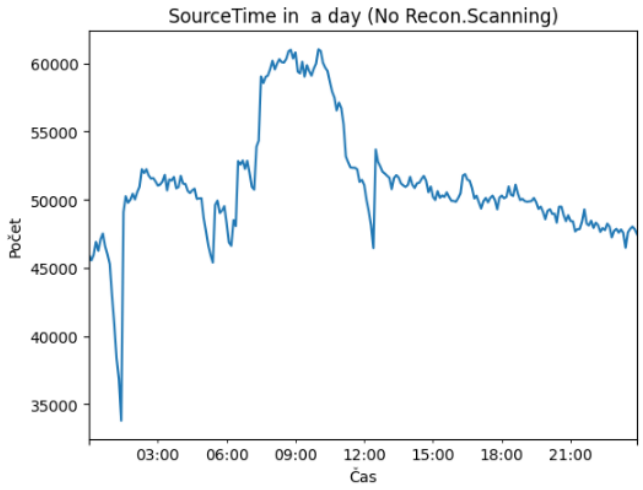


Obrázok 26: Tepelná mapa nad dňom v týždni a hodinou počas pracovného dňa (čas zdrojového zariadenia).

Z obrázka č. 26 a z výsledkov analýzy nad celou našou dátovou sadou môžeme tvrdiť, že útočníci podľa našej dátovej sady preferujú využívanie zariadení na útok primárne v dňoch pondelok a utorok pričom podľa našej dátovej sady sa útočníci vyhýbajú zneužitiu zariadení v dni štvrtok. Taktiež môžeme vidieť, že útočníci ak plánujú zneužiť zariadenie na útok, tak preferujú, aby sa dané zariadenie nachádzalo v pracovných hodinách mimo obedňajšej prestávky, pripadne v poobedňajších pracovných hodinách. To nám potvrdzuje vyššie získané poznatky. Na zariadeniach pravdepodobne sa spúšťajú automatizované činnosti a v prípadne obedňajších prestávok sa pozostavujú.

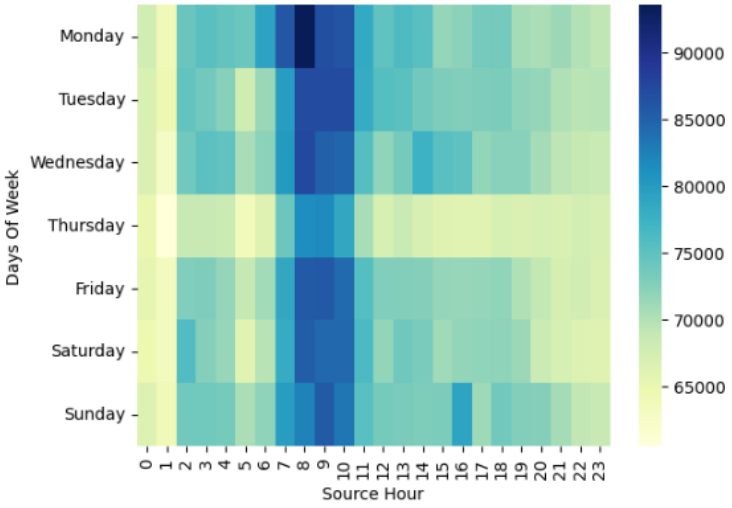
5.4.2 Rozdelenie bezpečnostných udalosti podľa kategórií udalostí

V predošlých častiach prace sme načrtli jednotlivé kategórie udalostí vyskytujúcich sa v systéme Warden. Z tohto dôvodu sme sa rozhodli pozrieť na našu dátovú sadu bez určitých kategórií. Rozhodli sme sa pozrieť na bezpečnostné udalosti, ktoré nie sú kategorizované ako Recon.Scanning, keďže Recon.Scanning je kategória z najväčším percentuálnym obsadením, až 94% z bezpečnostných udalostí bez testovacích dát. Následne sme vykonali analýzu nad dátovou sadou bez udalostí kategórií Recon.Scanning, Attempt.Login a Intrusion.UserCompromise, čo odpovedá odstráneniu 99.9% z našej očistenej dátovej sady, avšak výsledná dátová sada, aj keď je len 0.01%, stále reprezentuje 1.774.481 bezpečnostných udalostí.



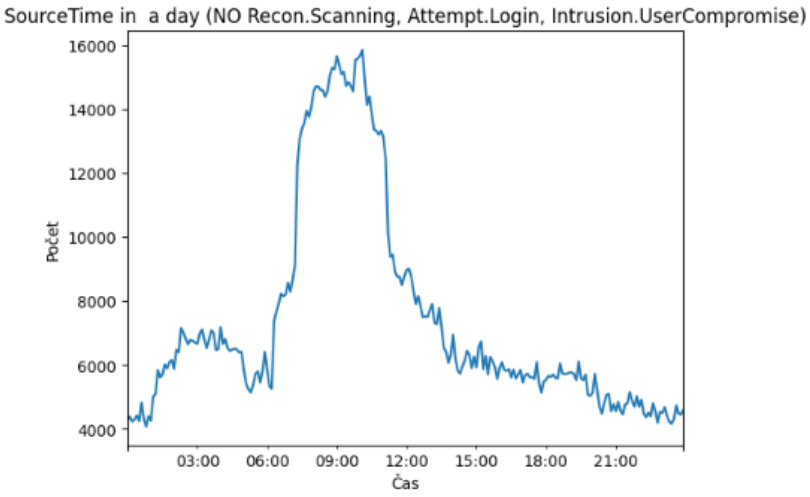
Obrázok 27: Časový rad počtu udalostí v dni bez kategórie Recon.Scanning (čas zdrojového zariadenia).

Na obrázku č. 27 vidíme výsledok prvej analýzy dátovej sady bez kategórie Recon.Scanning. Oproti analýze nad celou dátovou sadou sa táto líši relatívnou stabilitou s výnimkou v čase od 8:00 do 11:00, kde počet udalostí bol vyšší a krátkym poklesom okolo 1:00.



Obrázok 28: Tepelná mapa nad dňom v týždni a hodinou bez kategórie Recon.Scanning (čas zdrojového zariadenia).

Na tepelnej mape (Obr. č. 28) môžeme vidieť rovnaký trend vyššieho počtu udalostí v predobedňajších hodinách a pokles okolo 1:00, ako to bolo u analýzy z Obr. č. 26. Taktiež môžeme vidieť, že dochádza k systematickému poklesu útokov vo štvrtok, čo môžeme vidieť aj u analýzy nad celou dátovou sadou (Obr. č. 24).



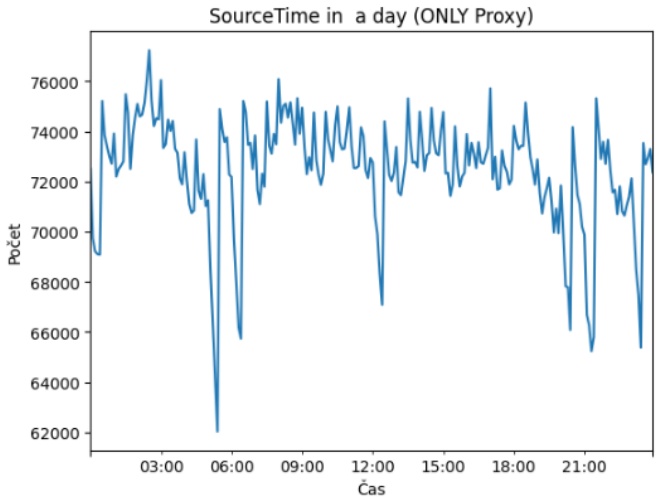
Obrázok 29: Časový rad počtu udalostí v dni bez kategórií Recon.Scanning, Attempt.Login, Intrusion.UserCompromise (čas zdrojového zariadenia).

Oproti analýze nad celou dátovou sadou v prípade analýzy z Obr. č. 29 vidíme prudký nárast útokov v časovom úseku od 7:00 do 12:00 spolu s malým nárastom útokov v čase od 1:00 do 5:00. Pomocou odfiltrovania najpočetnejších kategórii sme dostali len

špecifické kategórie, ktoré reprezentujú pokročilejšie útoky ako napríklad Malware.Trojan, Availability.DDoS atď. Z tejto analýzy teda vieme konštatovať, že podľa našej dátovej sady útočníci preferujú vykonávať zložitejšie útoky z zneužitých zariadení v dopoludňajších hodinách.

5.4.3 Rozdelenie bezpečnostných udalosti podľa typu zariadenia

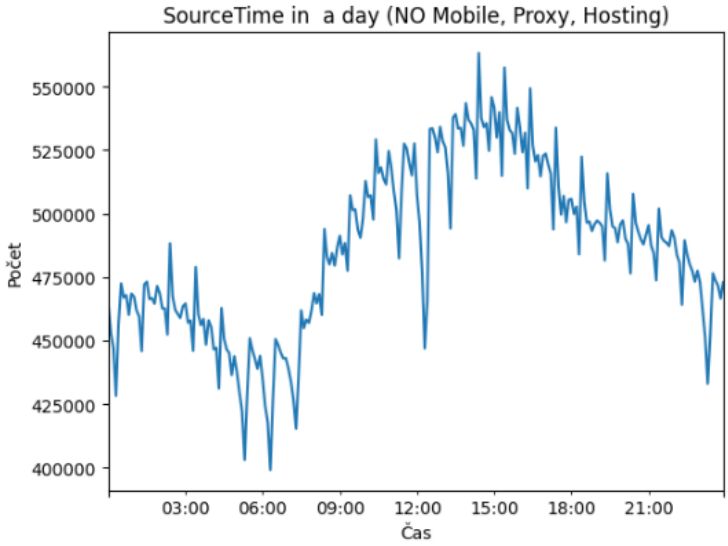
V tejto časti našej práce sme sa rozhodli zamerať na rozdelenie bezpečnostných udalostí podľa typu zariadenia. Jednotlivé typy pre dané zariadenia z našej dátovej sady sme získali pomocou služby IP-API [18]. Táto služba poskytuje parametre proxy, mobile a hosting, ktoré bližšie popisujú typ zariadenia nachádzajúci sa na danej IP adrese. IP adresy typu proxy sú zariadenia, ktoré slúžia ako VPN služby, TOR exit node a ďalšie. IP adresy typu mobile sú IP adresy, ktoré sú priradene mobilným operátorom a slúžia pre mobilne služby. V poslednom rade IP adresy typu hosting sú IP adresy priradene k dátovým centrám a hostingovým službám.



Obrázok 30: Časový rad počtu udalostí v dni pre typ zariadenia Proxy (čas zdrojového zariadenia).

V prípade zariadení typu Proxy môžeme z našej analýzy z Obr. č. 30 vidieť mierne poklesy útokov zo zneužitých zariadení v čase okolo 6:00 ráno, okolo 12:00 a približne okolo 21:00. Tieto poznatky priamo nadväzujú na analýzy z celej dátovej sady kde sme videli rovnaké správanie útočníkov pri výbere zariadení na útok.

V poslednom rade sme sa pozreli na zariadenia, ktoré nespadajú do kategórií mobile, proxy a hosting.



Obrázok 31: Časový rad počtu udalostí v dni bez typov zariadení Mobile, Proxy a Hosting (čas zdrojového zariadenia).

Pri tejto analýze (Obr. č. 31) môžeme vidieť relatívne rovnaký trend oproti celej dátovej sade. Môžeme vidieť pokles počtu útokov zo zariadení v časoch okolo 6:00 ráno a nárast v počte okolo 14:00 a následný pokles v nočných hodinách.

5.4.4 Rozdelenie bezpečnostných udalosti podľa reputácie zariadenia

Nakoniec sme sa zamerali na analýzy z pohľadu rozdelenia bezpečnostných udalostí podľa reputácie zariadenia. Či už podľa reputačného skóre, podľa výskytu v určitých blacklistoch prípadne podľa počtu blacklistov, v ktorých figuruje daná IP adresa. Ako už bolo spomínané pri predstavovaní služby NERD, naše analýzy budú vykonávane nad dátovou sadou z apríla 2024. Ide o čerstvé dáta, ktorými služba NERD disponuje.V prípade časti dáta z roku 2021 nie je možné vykonať takéto doplnenie údajov. Aj keby to bolo možné, životnosť týchto údajov je rádovo niekoľko týždňov, čo dáta z roku 2021 nespĺňajú.

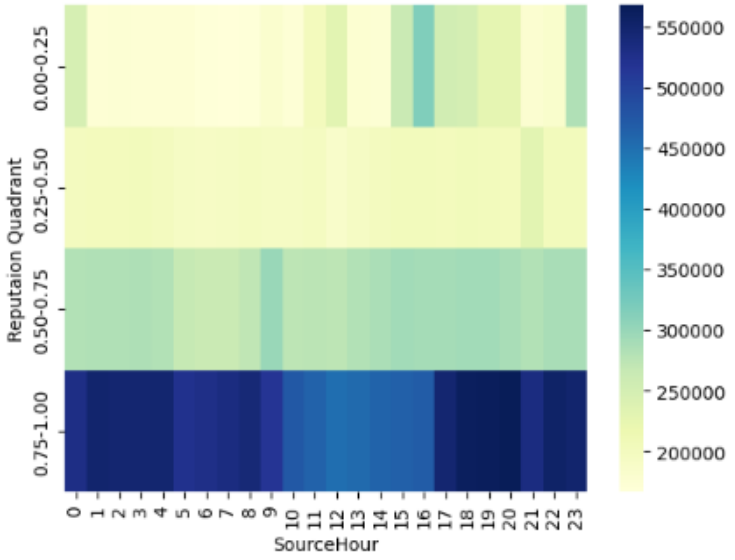
V prvom rade sme si rozdelili naše bezpečnostné udalosti do štyroch skupín. Tieto skupiny reprezentujú rozmedzie reputačných skóre. Do prvej skupiny spadajú reputačné

skóre v rozmedzí od 0.0 do 0.25 a tento trend pokračuje až do 1.00. Výsledne skupiny sú teda od 0.0 do 0.25, od 0.25 do 0.5, od 0.5 do 0.75 a nakoniec od 0.75 do 1.0.

Tabuľka 5: Najpočetnejšie skupiny reputačných skóre z bezpečnostných udalosti

Skupina	Počet
Od 0.75 do 1.00	12 496 198
Od 0.00 do 0.25	10 560 950
Od 0.50 do 0.75	6 785 201
Od 0.25 do 0.50	4 813 840

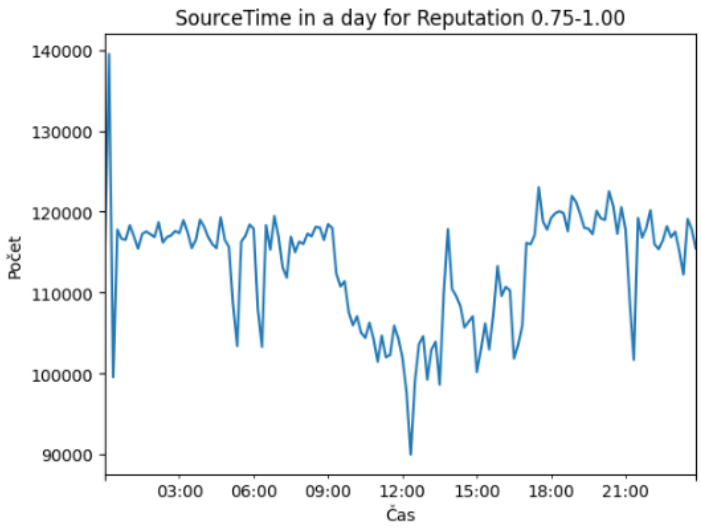
Z tabuľky č. 4 môžeme vidieť, že najpočetnejšou skupinou sú bezpečnostné udalosti, v rámci ktorých figuruje IP adresa s reputačným skóre od 0.75 do 1.00. To znamená, že v našej dátovej sade, IP adresy, ktoré majú najvyššiu tendenciu podieľať sa na útokoch sú aj pravé tie, ktoré sa aj najviac podieľajú na útokoch. Taktiež v našich analýzach nebudeme dbať na reputačné skóre 0.00, keďže dane číselná reprezentácia reprezentuje reputačné skóre pre neidentifikovane IP adresy pomocou služby NERD.



Obrázok 32: Tepelná mapa nad skupinami reputačného skóre a hodinou zdrojového zariadenia.

Z Obr. č. 32 môžeme potvrdiť, že skupina reputačných skóre od 0.75 do 1.00 je najpočetnejšia a taktiež môžeme pozorovať mierny pokles útokov z pohľadu zdrojových zariadení v čase okolo od 10:00 do 16:00. Bližším pozretím sa na udalosti, ktoré patria do skupiny od 0.00 do 0.25, vidíme mierny nárast udalostí v čase 00:00, 16:00 a o 23:00. V týchto časoch nám prudko narástol počet bezpečnostných udalostí kategórie Intrusion.UserCmpromise až na 27% v danom úseku. Bezpečnostné udalosti kategórie Malware zrástli oproti celku z pár desatín % na vyše 3%. Taktiež sa ukázalo, že najpočetnejšia krajina zdrojových zariadení z daných bezpečnostných udalostí bola Nemecko, a to až 18%, pričom pri kategórii Intrusion.UserCompromise to bolo až 55% z bezpečnostných udalosti tejto kategórie.

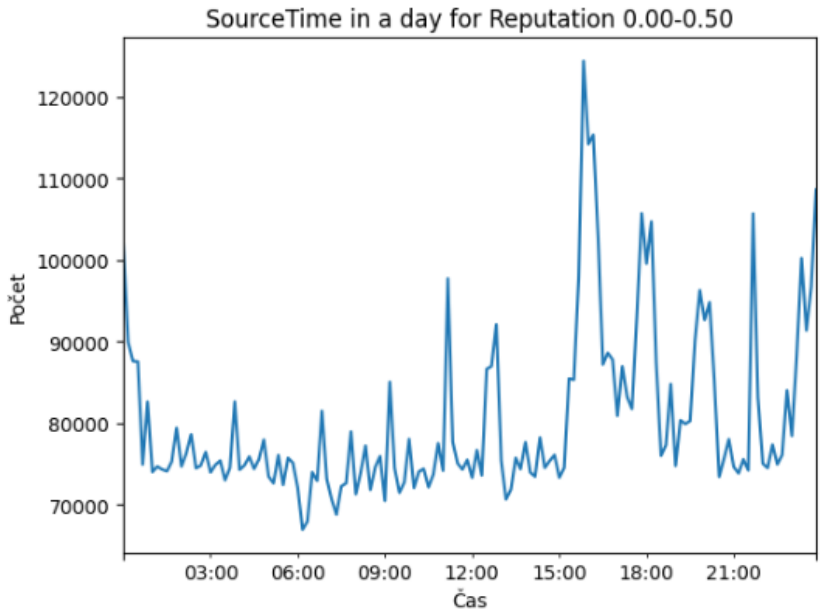
Keďže skupina s reputačným skóre od 0.75 do 1.00 je najpočetnejšia a súčasne vidíme pokles daných bezpečnostných udalostí v čase od 10:00 do 16:00, spravíme analýzy počtu bezpečnostných udalosti v dni práve pre túto skupinu.



Obrázok 33: Časový rad počtu udalostí v dni pre reputačné skóre od 0.75 do 1.00

Na Obr. č. 33 môžeme vidieť výsledok analýzy pre počet udalostí v dni z pohľadu zdrojových zariadení, ktoré sú v skupine reputačného skóre 0.75 a vyššie. Na časovom rade tejto analýzy môžeme vidieť jasný pokles útokov okolo 12:00 a potom o 21:00. Tento trend je podobný ako u predošlých analýz, kde sme mohli pozorovať podobné výsledky. V ďalšej časti sa pozrieme na skupiny reputačných skóre, ktoré sú menej

notorické k útokom, a teda s menším reputačným skóre. Do tejto skupiny patria zariadenia (IP adresy) s reputačné skóre 0.50 a menej.



Obrázok 34: Časový rad počtu udalostí v dni pre reputačné skóre od 0.00 do 0.50

Obr. č. 34 opisuje pravé skupinu IP adries s nízkym reputačným skóre, a teda nízkou tendenciou podieľania sa na útokoch. Čo môžeme vidieť v tomto časovom rade, je nárast útokov z pohľadu zneužitých zariadení v poobedňajších hodinách, pričom jednotlivé nárasty prichádzajú v separátnych vlnách.

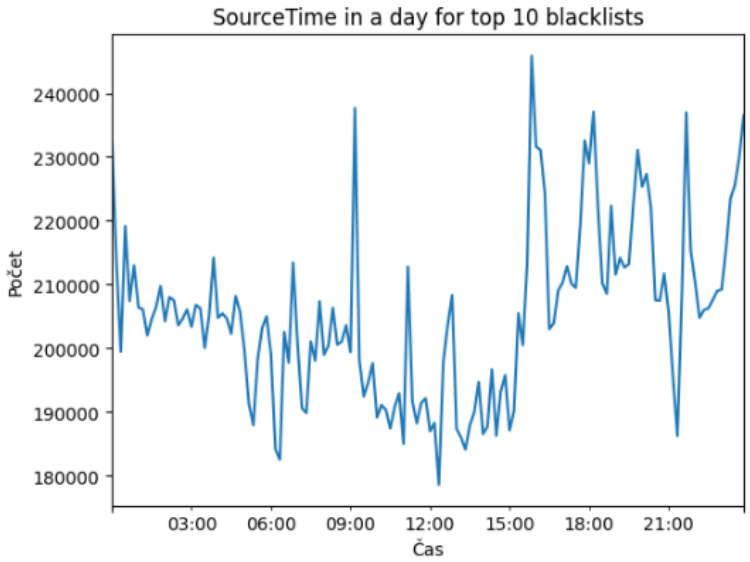
V poslednom rade sme sa zamerali na výskyt v blacklistoch. Celkovo sa v našej obohatenej dátovej sade vyskytujú 62 rôznych blacklistov. Je doležíte podotknúť, že daná IP adresa môže figurovať aj vo viacerých blacklistoch zároveň, a teda čím vyššiu reputáciu ma daná IP adresa (vyšší podiel na počte bezpečnostných udalosti), tým viacej bude daná IP adresa figurovať na rôznych blacklistoch. Najprv sme sa zamerali na to, ktoré blacklisty sú najpočetnejšie.

Tabuľka 6: Prvých 5 najpočetnejších skupin blacklistov z bezpečnostných udalosti

Blacklist	Počet
Abuseipdb	24 714 343
Turris_greylist	22 021 599

Ciarmy	18 488 810
dshield	9 035 610
Dataplane_org_sshclient	8 321 475

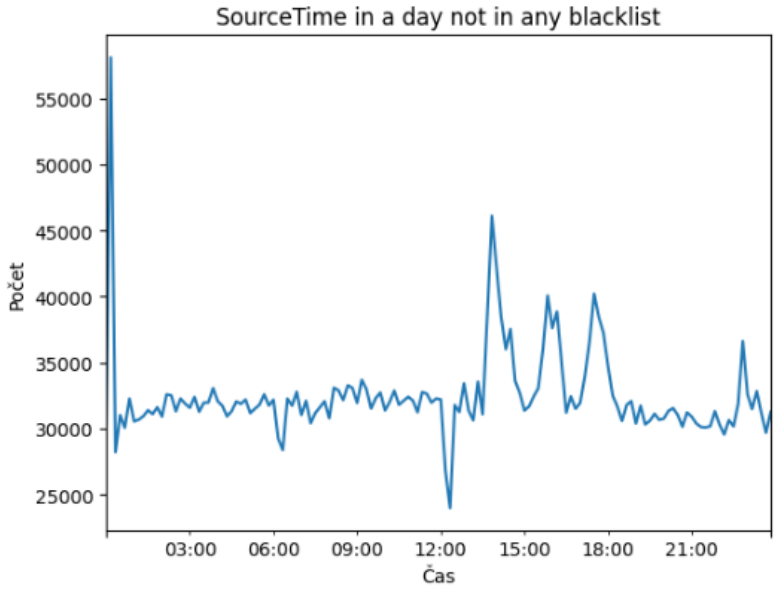
Z tabuľky č. 5 môžeme vidieť, že najdominantnejšie blacklisty sú abuseipdb, turris_greylist a ciarmy. Abuseipdb dokonca figuruje až v takmer 25 miliónov bezpečnostných udalostiach, čo odpovedá približne 71% bezpečnostných udalostí z apríla 2024. Podobne ako u predošlých analýz aj teraz spravíme analýzu pre počet udalostí v dni z pohľadu zdrojového zariadenia, avšak pre 10 najpočetnejších blacklistov, ktoré sa nachádzajú v našej obohatenej dátovej sade.



Obrázok 35: Časový rad počtu udalostí v dni pre 10 najpočetnejších blacklistov

Podobne ako to bolo u reputačných skóre, aj u analýzy z Obr.č. 35, kde sme analyzovali počet udalostí v dni pre 10 najpočetnejších blacklistov, môžeme vidieť nárast útokov z pohľadu zneužitých zariadení v poobedňajších hodinách. Súčasne jednotlivé nárasty, tak isto ako u analýzy z Obr. č. 34, prichádzajú v separátnych vlnách.

V poslednom rade sme sa zamerali na počet, v koľkých blacklistoch figurujú IP adresy z bezpečnostných udalostí v našej dátovej sade. Presnejšie sme sa zamerali na také IP adresy, ktoré nefigurujú v žiadnom z blacklistov. Predpokladáme teda, že dané IP adresy reprezentujú zneužité zariadenia bežných používateľov.



Obrázok 36: Časový rad počtu udalostí v dni pre bezpečnostné udalosti bez výskytu IP adresy v blackliste

Pri analýze, ktorú môžeme vidieť na Obr, č. 36 je jasný nárast bezpečnostných udalostí z pohľadu zneužitých zariadení v časoch 00:00 a poobedňajších hodinách okolo 13:00, 16:00 a 17:00 s jasným poklesom okolo 12:00. Pokles o 12:00 môžeme vidieť na viacero predošlých analýzach napr. na Obr. č. 32 a 22.

Záver

Stály rast zariadení pripojených v Internete ma za dôsledok zvýšený počet bezpečnostných upozornenia a udalostí. Tento zvýšený počet bezpečnostných upozornení a udalostí má za následok zvýšené nároky na bezpečnostných analytikov. Jedným zo spôsobov riešenia tohto problému je implementácia nových metód a prístupov pre analýzu bezpečnostných udalostí v kybernetickej bezpečnosti. Medzi prospešné prístupy sa javí aj časopriestorová analýza, ktorej je venovaná aj táto práca.

Prvým cieľom našej bakalárskej práce bolo spracovanie, úprava a obohatenie dátovej sady na vykonanie časopriestorovej analýzy. Tomuto cieľu sme venovali podstatnú časť nasej práce, keďže sa o komplexné upravovanie veľkej dátovej sady (681 GB), pri ktorom treba zhodnotiť výpočtovú a pamäťovú zložitosť programov. Tomuto cieľu sme sa venovali v tretej a štvrtej kapitole. V tretej kapitole sme vysvetlili jednotlivé externé služby, ako NERD a IP-API, potrebné pre našu analýzu z hľadiska obohatenia dátovej sady. V štvrtej kapitole sme načrtli schému celého nášho procesu spolu s postupnými krokmi, ktoré sme vykonali na spracovanie, obohatenie a úpravu dátovej sady. Medzi tieto kroky patrí aj využitie spomínaných externých služieb ako IP-API a NERD na obohatenie bezpečnostných udalostí o geolokačný a reputačný parameter zdrojového zariadenia. Následne sa v štvrtej kapitole venujeme pridaniu času zdrojového zariadenia do dátovej sady pomocou dát z externých služieb a finálnym úpravám sady o atribúty, ktoré upravujú dátovú sadu do podoby, s ktorou môžeme úspešne a rýchlo vykonať časopriestorovú analýzu.

Druhým cieľom v našej práci bolo porovnanie aktuálnych prístupov v časopriestorovej analýzy. Tento cieľ sme zhrnuli v prvej a poslednej kapitole. V prvej kapitole sme stručne popísali potrebné veci na našu časopriestorovú analýzu spolu z popismi základných pojmov. Ako ďalšie sme v prevej kapitole popísali jednotlivé prístupy časopriestorových analýz použitých v podobných prácach. V poslednej kapitole sme sa definovali prístupy vizualizácie analýz ako napríklad časové rady a tepelné mapy, spolu s výhodami jednotlivých prístupov pri interpretácii výsledkov časopriestorovej analýzy v kybernetickej bezpečnosti.

Ako posledný cieľ našej bakalárskej práce bolo analyzovať dátovú sadu pomocou vybraných prístupov časopriestorovej analýzy a následne vyhodnotenie a interpretácia týchto výsledkov. Tento cieľ sme popísali v poslednej kapitole, kde sme primárne

vysvetľovali rôzne interpretácie našich výsledkov. Medzi vybrané prístupy patrili vizualizácie pomocou časových radov a tepelných máp. Naše analýzy boli zamerané na významnosť a rolu času zdrojového zariadenia v útoku. Toto zameranie bolo odzrkadlené na počte analýz zameraných na čas zdrojového zariadenia. Taktiež sme vykonali jednotlivé analýzy aj z iných pohľadov ako napríklad za pomoci reputačného ohodnotenia IP adresy a mnoho ďalších.

Zoznam použitej literatúry

1. HAINING, Robert P. Spatial data analysis: theory and practice. Cambridge university press, 2003.
2. SHEKHAR, Shashi; XIONG, Hui (ed.). Encyclopedia of GIS. Springer Science & Business Media, 2007.
3. ISO 22300:2021 [online] [cit. 2024-05-17] Dostupné z https://www.iso.org/standard/77008.html
4. Defining a Security Incident vs Event: When to Report [online] [cit. 2024-05-15] Dostupné z https://www.iansresearch.com/resources/all-blogs/post/security-blog/2023/09/05/defining-a-security-incident-vs-event-when-to-report
5. Amin, R. W., Sevil, H. E., Kocak, S., Francia III, G., & Hoover, P. (2020). The spatial analysis of the malicious uniform resource locators (URLs): 2016 dataset case study. Information, 12(1), 2.
6. Sethi, A. A. R. U. S. H. I. "Statistical Methods for Spatial Data Analysis." Cyber Secur. Insights Mag 1 (2022): 7-11.
7. Sokol, Pavol, and Veronika Kopčová. "Lessons learned from correlation of honeypots' data and spatial data." 2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE, 2016.
8. P. Sokol, L. Kleinová and M. Husák, "Study of attack using honeypots and honeynets lessons learned from time-oriented visualization," IEEE EUROCON 2015 - International Conference on Computer as a Tool (EUROCON), Salamanca, Spain, 2015, pp. 1-6, doi: 10.1109/EUROCON.2015.7313713.
9. ZUZČÁK, Matej; BUJOK, Petr. Causal analysis of attacks against honeypots base on properties of countries. IET Information Security, 2019, 13.5: 435-447.
10. CARRIEGOS VIEIRA, Miguel, et al. Towards Supercomputing Categorizing the Maliciousness upon Cybersecurity Blacklists with Concept Drift. Computational and Mathematical Methods, 2023, 2023: 1-8.
11. ČERGEŤ, Maroš; HUDEC, Ján. Cyber-Security Threats Origins and their Analysis. Acta Polytechnica Hungarica, 2023, 20.9.
12. FAN, Zhijie, et al. An improved integrated prediction method of cyber security situation based on spatial-time analysis. Journal of Internet Technology, 2018, 19.6: 1789-1800
13. Warden. [online] Dostupné z https://warden.cesnet.cz/en/architecture

14. SIMOES, Paulo, et al. On the use of honeypots for detecting cyber attacks on industrial control networks. In: Proc. 12th Eur. Conf. Inform. Warfare Secur. ECIW. 2013
15. IDEA [online] Dostupné z https://idea.cesnet.cz/en/index
16. MUIR, James A.; OORSCHOT, Paul C. Van. Internet geolocation: Evasion and counterevasion. Acm computing surveys (csur), 2009, 42.1: 1-23.
17. IP2Location [online] [cit. 2024-02-03] Dostupné z https://www.ip2location.com/
18. IP-API [online] [cit. 2024-02-03] Dostupné z https://IP-API.com/
19. DB-IP [online] [cit. 2024-02-03] Dostupné z https://db-ip.com/
20. KESTER, Jan-Jelle. Comparing the accuracy of IPv4 and IPv6 geolocation databases. Methodology, 2016, 10.11: 12-17.
21. Nur, Abdullah Yasin. Accuracy and Coverage Analysis of IP Geolocation Databases. In: 2023 International Balkan Conference on Communications and Networking (BalkanCom). IEEE, 2023. p. 1-6.
22. MaxMind [online] [cit. 2024-04-14] Dostupné z https://www.maxmind.com/en/geoip-databases
23. Evaluating Reputation of Internet Entities [online] [cit. 2024-05-17] Dostupné z https://inria.hal.science/hal-01632738/file/385745_1_En_13_Chapter.pdf
24. Blacklist [online] [cit. 2024-05-17] Dostupné z https://csrc.nist.gov/glossary/term/blacklist
25. NERD Architektúra [online] [cit. 2024-05-15] Dostupné z https://github.com/CESNET/NERD/wiki/Architecture
26. NERD Reputačné skóre [online] [cit. 2024-05-15] Dostupné z https://github.com/CESNET/NERD/wiki/Reputation-score
27. Python 3.12.1 [online] [cit. 2023-02-03] Dostupné z https://www.python.org/downloads/release/python-3121/
28. VANDERPLAS, Jake. Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.", 2016.
29. Modul sqlite3 [online] [cit. 2024-02-03] Dostupné z: https://docs.python.org/3/library/sqlite3.html
30. Modul ipadress [online] [cit. 2024-02-03] Dostupné z: https://docs.python.org/3/library/ipaddress.html
31. Modul datetime [online] [cit. 2024-02-03] Dostupné z: https://docs.python.org/3/library/datetime.html

32. Modul pytz [online] [cit. 2024-02-03] Dostupné z: https://pypi.org/project/pytz/
33. Modul pandas [online] [cit. 2024-02-03] Dostupné z: https://pandas.pydata.org/docs/
34. Modul requests [online] [cit. 2024-02-03] Dostupné z: https://requests.readthedocs.io/en/latest/https://docs.python.org/3/library/sqlite3.html
35. Modul json [online] [cit. 2024-02-03] Dostupné z: https://docs.python.org/3/library/json.html
36. Modul matplotlib [online] [cit. 2024-02-03] Dostupné z: https://matplotlib.org/stable/index.htmlhttps://docs.python.org/3/library/json.html
37. Modul seaborn [online] [cit. 2024-02-03] Dostupné z: https://seaborn.pydata.org/index.html
38. Jupyter Notebook [online] [cit. 2024-02-03] Dostupné z: https://docs.jupyter.org/en/latest/
39. BROCKWELL, Peter J.; DAVIS, Richard A. Time series: theory and methods. Springer science & business media, 1991.
40. GEHLENBORG, Nils; WONG, Bang. Heat maps. Nature Methods, 2012, 9.3: 213.

Prílohy

Príloha A: Bakalárska práca v elektronickej podobe

Príloha B: Zdrojové kódy na spracovanie, obohatenie, finálne úpravy a analýzu dátovej sady