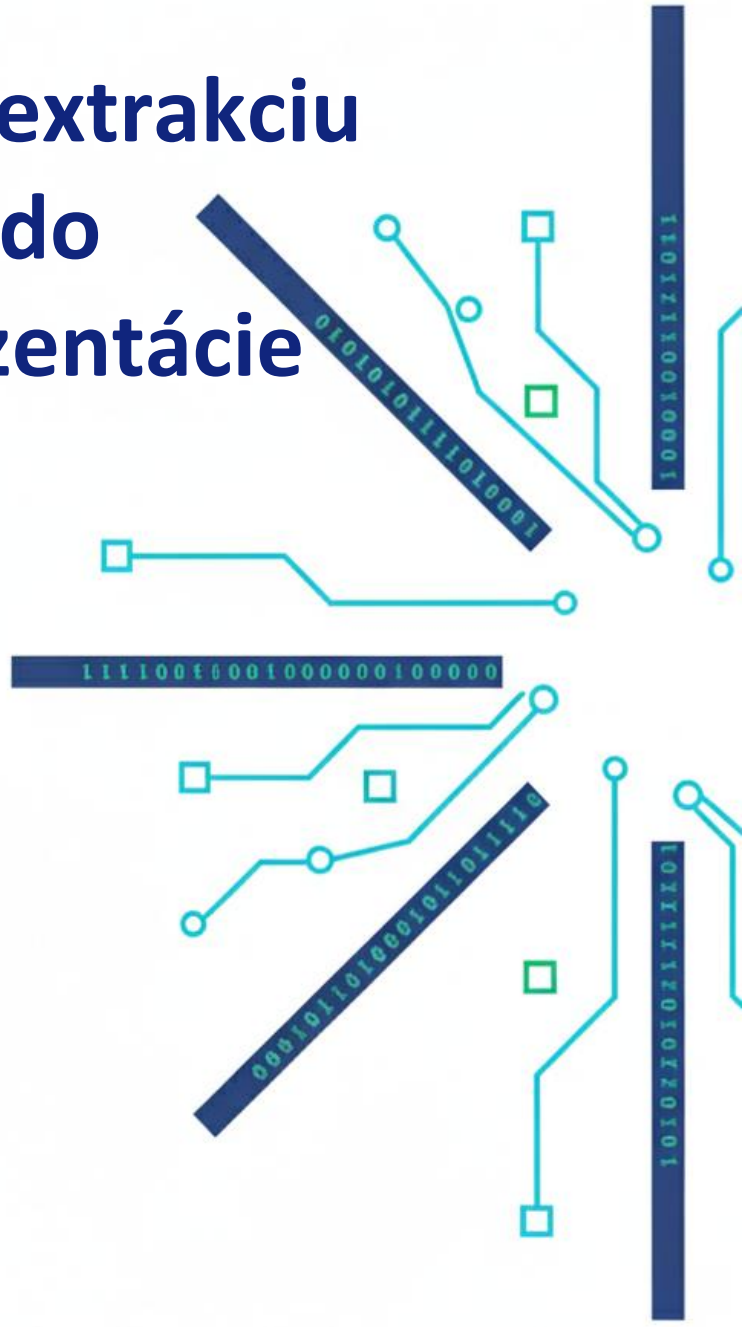
D15 – Model na extrakciu digitálnych stôp do maticovej reprezentácie



Projekt Automatizácia digitálnej forenznej analýzy a reakcie na incident (ADFIR) financovaný Európskou úniou – Next GenerationEU prostredníctvom Plánu obnovy a odolnosti Slovenskej republiky pod číslom projektu č. 09-I05-03-V02-00079.

Obsah

1 Popis projektu

Projekt **Automatizácia digitálnej forenznej analýzy a odpovede na incident** (ďalej len „**ADFIR**“) je financovaný **Európskou úniou – Next GenerationEU prostredníctvom Plánu obnovy a odolnosti Slovenskej republiky** pod číslom projektu č. 09-I05-03-V02-00079. Tento projekt sa zaoberá jednou z kľúčových výziev v oblasti kybernetickej bezpečnosti a informačnej bezpečnosti – ako spracovať obrovské množstvo digitálnych dôkazov, ktoré vznikajú počas incidentov kybernetickej bezpečnosti alebo forenzných vyšetrovaní. V súčasnosti je tento proces veľmi náročný z hľadiska ľudských zdrojov a času. Automatizácia pomocou metód strojového učenia môže preto výrazne **zlepšiť kvalitu digitálnej forenznej analýzy** a skrátiť čas potrebný na jej vykonanie. Celkovo to umožňuje bezpečnostným tímom efektívnejšie reagovať na kybernetické hrozby. Hlavné prínosy tohto projektu sú:

- **Rýchlejšie riešenie incidentov v oblasti kybernetickej bezpečnosti.** Projekt ADFIR zavádza automatizované prístupy k zberu, spracovaniu a analýze digitálnych stôp. Vďaka tomu môžu bezpečnostné tímy rýchlejšie identifikovať príčiny incidentov a prijať účinné opatrenia na ich riešenie.
- **Zníženie pracovnej záťaže forenzných analytikov**. Rutinné a časovo náročné úlohy spojené so spracovaním digitálnych stôp budú nahradené automatizovanými metódami. To umožní analytikom sústrediť sa na zložitejšie prípady a strategické rozhodovanie.
- **Vyššia kvalita a konzistentnosť výstupov.** Použitie jednotných metodík a nástrojov zaručuje, že spracované digitálne stopy budú presnejšie, konzistentnejšie a ľahšie overiteľné. To výrazne znižuje riziko chýb spôsobených ľudskými faktormi.
- **Možné využitie v trestnom konaní**. Výstupy projektu budú vyvinuté v súlade s právnymi požiadavkami a normami, čo umožní, aby digitálne stopy boli akceptované ako relevantné dôkazy pre vyšetrovanie a súdne konania.

2 Úvod

V oblasti digitálnej forenznej analýzy predstavuje spracovanie veľkého množstva heterogénnych dát zásadnú výzvu, najmä pokiaľ ide o ich ďalšie využitie v automatizovaných analytických metódach. Supertimeline, ako výstup nástrojov typu Plaso, poskytuje síce komplexný a informačne bohatý pohľad na priebeh udalostí v systéme, avšak jeho štruktúra nie je priamo vhodná pre aplikáciu metód strojového učenia alebo formálnych analytických prístupov.

Tento dokument sa zameriava na návrh modelu extrakcie digitálnych stôp do maticovej reprezentácie, ktorá umožní transformovať pôvodné, prevažne neštruktúrované alebo textové dáta do podoby vhodnej na ďalšie spracovanie. Cieľom je systematicky identifikovať relevantné atribúty, zakódovať ich do binárnych, kategorických alebo numerických premenných a zároveň minimalizovať stratu informácie pri tejto transformácii.

Navrhovaný prístup vychádza z potreby preklenúť medzeru medzi manuálnou forenznou analýzou a automatizovanými metódami spracovania dát. Kým tradičné prístupy sa sústreďujú najmä na interpretáciu jednotlivých artefaktov v časovej osi, cieľom tohto modelu je vytvoriť reprezentáciu, ktorá umožní efektívne využitie pokročilých analytických techník, ako sú metódy strojového učenia či redukcia dimenzie.

Tento výstup zároveň tvorí základ pre nadväzujúce spracovanie dát v rámci projektu ADFIR, konkrétne pre agregáciu a prepojenie digitálnych stôp, kde už transformované dáta umožňujú vyššiu úroveň abstrakcie, korelácie a interpretácie udalostí.

3 Proces spracovania digitálnych stôp

Vstupným podkladom pre navrhovaný model na extrakciu digitálnych stôp do maticovej reprezentácie boli dáta získané prostredníctvom nástroja Plaso [7]. Tento nástroj umožňuje agregáciu a koreláciu forenzných artefaktov z rôznych zdrojov do jednotnej časovej osi, tzv. supertimeline. Výsledná dátová štruktúra má charakter tabuľky pozostávajúcej zo 17 atribútov, ktoré sú bližšie špecifikované v Tabuľke 1.

Atribút	Popis	Typ
Date	dátum, kedy nastala udalosť	object
Time	čas, kedy nastala udalosť	object
Timezone	časová zóna	object
MACB	časové pečiatky (Modification, Access, Creation, Birth)	object
Source	skratka názvu zdroja (napr. REG - záznamy z registra)	object
Sourcetype	popis zdroja	object
Type	typ časovej značky (napr. posledný zápis)	object
User	meno používateľa (ak existuje), ktoré je spojené s udalosťou	object
Host	názov hostiteľa (ak existuje), ktorý je priradený k udalosti	object
Short	obsahuje pole s krátkym popisom, v ktorom je uložený text	object
Desc	pole, ktoré obsahuje väčšinu analyzovaných informácií	object
Version	číslo verzie časovej pečiatky	int64
Filename	názov súboru, ktorý je spojený s udalosťou	object
Inode	číslo inodu analyzovaného súboru	object
Notes	miesto na ukladanie dodatočných informácií	object
Format	vstupný modul, ktorý bol použitý na analyzovanie	object
Extra	pole s parsovanými informáciami, ktoré sú tu spojené a uložené	object

Table 1 - Popis atribútov supertimeline

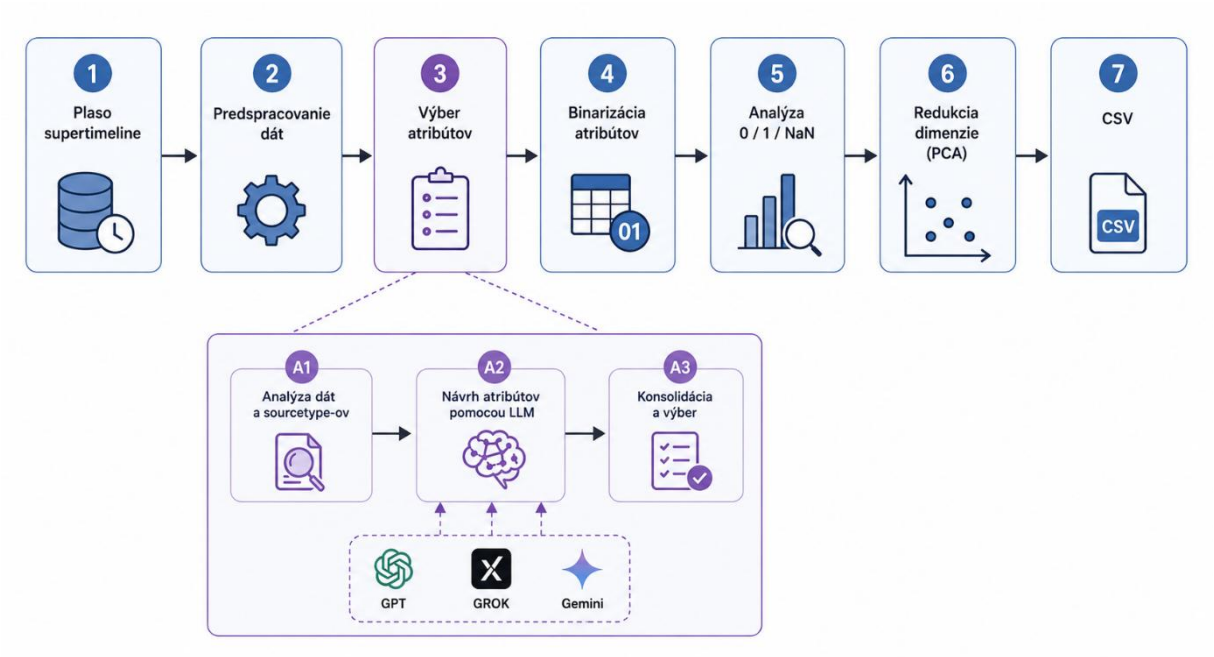
Supertimeline predstavuje komplexný a informačne bohatý zdroj údajov, ktorý je primárne určený pre potreby manuálnej digitálnej forenznej analýzy. V tomto kontexte umožňuje analytikovi rekonštruovať časovú postupnosť udalostí a identifikovať relevantné aktivity v skúmanom systéme. Napriek vysokej výpovednej hodnote však táto forma dát nie je priamo

vhodná na aplikáciu automatizovaných analytických metód, akými sú napríklad metódy strojového učenia, formálna konceptová analýza či prístupy založené na teórii grafov.

Hlavným problémom je predovšetkým dátová reprezentácia jednotlivých atribútov. S výnimkou atribútu Version, ktorý je reprezentovaný numerickým typom (int64), sú všetky ostatné atribúty uložené ako dátový typ object. V praxi to znamená, že ide prevažne o textové údaje, často vo forme neštruktúrovaných alebo len čiastočne štruktúrovaných reťazcov. Takáto heterogénna a semanticky bohatá reprezentácia síce poskytuje flexibilitu pri manuálnej interpretácii, avšak výrazne komplikuje ich ďalšie spracovanie v kontexte formálnych a kvantitatívnych analytických metód.

Transformácia týchto atribútov do vhodnej podoby predstavuje netriviálny problém, keďže si vyžaduje identifikáciu relevantných znakov, extrakciu významovej informácie a jej následné zakódovanie do štruktúrovanej reprezentácie. Konkrétne ide o prevod dát do binárnych, kategorických alebo numerických premenných, ktoré sú kompatibilné s požiadavkami algoritmov strojového učenia a ďalších analytických nástrojov. Pri tomto procese je zároveň nevyhnutné minimalizovať stratu informácie, ktorá by mohla negatívne ovplyvniť kvalitu následnej analýzy.

Cieľom tohto výstupu je preto navrhnúť systematický prístup k predspracovaniu dát zo supertimeline, ktorý umožní ich **transformáciu do maticovej reprezentácie** vhodnej pre ďalšie spracovanie. Dôraz je kladený najmä na zachovanie čo najväčšieho množstva relevantnej informácie, redukciu redundancie a zabezpečenie kompatibility s vybranými analytickými metódami. Výsledkom by mala byť dátová reprezentácia, ktorá umožní efektívne využitie pokročilých techník analýzy dát pri skúmaní digitálnych stôp.



Obrázok 1 – Proces spracovania digitálnych stôp

Na Obrázku 1 je znázornený proces spracovania digitálnych stôp od ich získania až po vytvorenie finálneho datasetu vhodného pre ďalšiu analýzu. Pipeline začína vstupnými dátami vo forme Plaso supertimeline, ktoré prechádzajú fázou predspracovania. Následne nasleduje výber a návrh atribútov, ktorý predstavuje kľúčový krok celého procesu.

Tento krok je detailnejšie rozpracovaný v spodnej časti diagramu. Zahŕňa analýzu dát a dostupných sourcetype-ov, generovanie kandidátnych atribútov pomocou veľkých jazykových modelov (GPT, Grok, Gemini) a následnú konsolidáciu a výber finálneho zoznamu atribútov.

Po definovaní atribútov nasleduje ich extrakcia a transformácia do binárnej reprezentácie (0/1/NaN), pričom NaN hodnoty reprezentujú nerelevantnosť atribútu pre daný záznam. Následne je vykonaná analýza distribúcie hodnôt a redukcia dimenzie pomocou metódy PCA. Výsledkom celého procesu je finálny dataset vo formáte CSV, pripravený pre ďalšie spracovanie, napríklad agregáciu alebo aplikáciu metód strojového učenia.

4 Výber atribútov

Výber atribútov predstavuje kľúčovú fázu transformácie forenzných dát do vhodnej maticovej reprezentácie, ktorá umožňuje ich ďalšie analytické spracovanie. Cieľom tejto fázy je identifikovať také atribúty, ktoré najlepšie vystihujú charakter skúmaných artefaktov a zároveň zachovávajú relevantné informácie pre detekciu incidentov. Proces výberu atribútov bol realizovaný viacstupňovo, pričom kombinoval štatistickú analýzu dát, využitie veľkých jazykových modelov a následnú konsolidáciu výsledkov. Dôležitým aspektom bolo zabezpečenie rovnováhy medzi komplexnosťou reprezentácie a jej interpretovateľnosťou. Osobitná pozornosť bola venovaná eliminácii redundantných a málo informatívnych atribútov.

4.1 Analýza dát a typov zdrojov

V úvodnej fáze spracovania dát sme sa zamerali na analýzu distribúcie jednotlivých artefaktov v rámci skúmaných prípadov. Na tento účel boli vytvorené základné kvantitatívne štatistiky, ktoré umožnili identifikovať najčastejšie sa vyskytujúce typy záznamov a zdrojov údajov. Táto analýza poskytla prehľad o dominantných artefaktoch v dátach a zároveň slúžila ako podklad pre ďalšie kroky návrhu reprezentácie.

Konkrétne výsledky analýzy sú uvedené v prílohe k tomuto výstupu - „D15 - Model na extrakciu digitálnych stôp do maticovej reprezentácie – výsledky“ v časti „sourcetype_source_counts_summar“.

Analyzovaná sumarizačná tabuľka *sourcetype_source_counts_summar* slúži na porovnanie výskytu jednotlivých **sourcetype (typ zdroja)** naprieč viacerými dátovými sadami. Každý **riadok** reprezentuje jeden konkrétny sourcetype, zatiaľ čo jednotlivé **stĺpce** predstavujú samostatné dátové sady alebo scenáre. **Hodnota v bunke** vyjadruje počet výskytov daného sourcetype v konkrétnej dátovej sade. Farebné rozlíšenie slúži na rýchlu interpretáciu vhodnosti:

- **zelená farba** označuje **vhodné typy zdrojov** (sourcetype) pre ďalšie spracovanie a maticovú reprezentáciu,
- **červená farba** označuje **nevhodné typy zdrojov** (sourcetype) pre ďalšie spracovanie a maticovú reprezentáciu,
- **oranžová farba** označuje **hraničné typy zdrojov** (sourcetype) viazané na špecifické prostredie (napr. konkrétny operačný systém).

Metodika hodnotenia vychádza z kombinácie frekvencie výskytu a konzistencie naprieč datasetmi. Dôležitým kritériom bolo aj zabezpečenie reprezentatívnosti atribútov pre rôzne typy forenzných scenárov.

4.1.1 Analýza vhodných typov zdrojov

Typy zdrojov (sourcetype), ktoré sú vhodné pre ďalšie spracovanie a maticovú reprezentáciu, môžeme rozdeliť do dvoch skupín, ktoré sa líšia počtom výskytu.
Typy zdrojov (sourcetype) s vysokou frekvenciou výskytu vo všetkých dátových sadách. Tieto zdroje sú vhodné, pretože sú konzistentne prítomné, bohaté na informácie a majú vysokú hodnotu pre forenzné vyšetrovanie. Medzi najvýznamnejšie patria tie zdroje, ktoré dosahujú maximálne alebo takmer maximálne hodnoty vo všetkých dátových sadách:

- **FILE | File entry shell item, FILE | File stat, FILE | NTFS USN change, FILE | NTFS file stat** – metadáta súborového systému s vysokou dostupnosťou a významným obsahom časových a systémových informácií,
- **EVT | WinEVTX** – záznamy (logy) udalostí, ktoré predstavujú kľúčový zdroj pre rekonštrukciu udalostí,
- **REG | Registry Key** – záznamy v registri operačného systému Windows, ktoré predstavujú generické záznamy registra bez bližšej kategorizácie alebo špecifického kontextu.

Druhým typom zdrojov (sourcetype) sú zdroje s nižšou frekvenciou výskytu, ale vysokou konzistenciou. Ich význam pre forenzné vyšetrovanie spočíva v tom, že poskytujú špecifické, ale opakovane využiteľné informácie. Ide o tie zdroje, ktoré sa síce nevyskytujú vo veľkom množstve, ale sú prítomné vo väčšine dátovych sád:

- **WEBHIST artefakty** (**WEBHIST | MSIE WebCache (container, cookies, records), WEBHIST | Chrome (Cache, Cookies, History)**) – dôležité pre analýzu používateľskej aktivity,
- **REG | Task Cache Registry Key** – artefakty naplánovaných úloh relevantné pre perzistenciu útokov,
- **REG | Registry Key – Service, REG | Service/Driver Configuration Registry Key, REG | BagMRU Registry Key**) – záznamy v registri operačného systému Windows, ktoré poskytujú štruktúrované a kontextovo bohaté informácie viazané na konkrétnu oblasť systému, ako je konfigurácia služieb a ovládačov (perzistencia, štart systému) alebo používateľská aktivita (napr. BagMRU – navigácia v súborovom systéme). Tieto artefakty majú vyššiu forenznú hodnotu, keďže umožňujú presnejšiu interpretáciu správania systému alebo používateľa,
- **LNK | Windows Shortcut** – poskytujú informácie o prístupe používateľa k súborom a aplikáciám, pričom sa vyskytujú naprieč viacerými dátovými sadami s relatívne stabilnou frekvenciou,

- **LOG | WinPrefetch** – napriek tomu, že sa nevyskytuje vo všetkých dátových sadách (najmä chýba v niektorých serverových scenároch), ide o významný zdroj pre analýzu spúšťania aplikácií,
- **OLECF | OLECF Item / Dest list entry** – reprezentujú artefakty súvisiace s prácou používateľa s dokumentmi (napr. recent files), pričom sa objavujú vo viacerých dátových sadách a poskytujú doplňujúce informácie k používateľskej aktivite.

4.1.2 Analýza nevhodných typov zdrojov

Tieto zdroje sú charakteristické veľmi nízkou frekvenciou výskytu, výraznou nekonzistentnosťou naprieč dátovými sadami alebo obmedzenou interpretačnou hodnotou. V mnohých prípadoch ide o artefakty (sourcetype), ktoré sa objavujú iba v jednom konkrétnom scenári alebo v špecifickej konfigurácii systému, čo výrazne znižuje ich využiteľnosť pre všeobecnú maticovú reprezentáciu.
Typickým príkladom sú artefakty ako **AMCACHEPROGRAM | Amcache Programs Registry Entry**, ktoré sa vyskytujú len v minimálnom počte datasetov, alebo **REG | Registry Key – RDP Connection**, ktorý je prítomný len v 1 dátovej sade. Podobne aj **PLIST | Plist file** reprezentuje platformovo špecifický artefakt, pričom sa v analyzovaných dátach objavuje iba okrajovo.

Ďalšiu skupinu tvoria logy a metadátové artefakty s obmedzeným kontextom, ako napríklad **LOG | Google Drive Sync Log** alebo **META | Open XML Metadata**, ktoré sa síce môžu v niektorých dátových sadách objaviť vo vyššom počte, avšak ich výskyt je striktne viazaný na konkrétnu aplikáciu alebo scenár (napr. cloud synchronizácia alebo práca s dokumentmi). Podobne artefakty typu **OLECF | OLECF Document Summary Info a OLECF | OLECF Summary Info** poskytujú iba doplnkové metadátové informácie bez priamej väzby na bezpečnostne relevantné udalosti, čo znižuje ich význam pri detekcii kybernetických bezpečnostných incidentov.

Špecifickú kategóriu predstavujú artefakty súvisiace s binárnymi súbormi, ako napr. **PE | PE Compilation time** alebo **PE | PE Import Time**, ktoré sú síce informatívne z pohľadu analýzy škodlivého kódu, ale vykazujú vysokú mieru nevyváženosti – v niektorých dátových sadách úplne absentujú, zatiaľ čo inde dosahujú vysoké hodnoty.

Medzi nevhodné patria aj artefakty s minimálnym výskytom, ako napríklad **RECBIN | Recycle Bin**, ktoré síce môžu obsahovať relevantné informácie, ale ich sporadický výskyt (jednotky až desiatky záznamov) neumožňuje ich efektívne využitie. Rovnako aj **WEBHIST | Chrome Autofill** alebo **WEBHIST | MSIE WebCache partitions record** sú silne závislé od konkrétneho používateľského správania a typu webového prehliadača.

Celkovo možno konštatovať, že tieto zdroje neprispievajú k robustnej a konzistentnej reprezentácii dát. Ich zahrnutie by viedlo k zvýšeniu šumu a potenciálnemu skresleniu

analytických modelov. Z tohto dôvodu boli tieto artefakty (sourcetype) v procese výberu atribútov vyradené a neboli zahrnuté do finálnej maticovej reprezentácie.

4.1.3 Analýza hraničných typov zdrojov

AMCACHE artefakty

Artefakty (sourcetype) typu AMCACHE (**AMCACHE | Amcache Registry Entry**) predstavujú významný zdroj informácií o spúšťaných aplikáciách v prostredí operačného systému Windows. Obsahujú záznamy o exe súboroch vrátane ich ciest, digitálnych odtlačkov (hashov) a časových údajov. Ich hlavnou výhodou je schopnosť zachytiť historické spúšťanie aplikácií, aj keď už samotné súbory nemusia byť prítomné v systéme. V analyzovaných datasetoch sa však nevyskytujú konzistentne vo všetkých prípadoch, čo súvisí s rozdielmi vo verziách operačného systému a spôsoboch zberu dát. V niektorých scenároch môžu byť tieto artefakty nedostupné alebo neúplné. Napriek tomu poskytujú vysokú forenznú hodnotu najmä pri identifikácii škodlivého softvéru. Ich využitie je vhodné najmä pri analýze kompromitovaných systémov. Z metodologického hľadiska sú vhodné ako doplnkový zdroj informácií. Ich nižšia univerzálnosť však obmedzuje ich použitie v globálnych modeloch. Preto sú zaradené medzi hraničné artefakty.

JOB (Scheduled Tasks) artefakty

Artefakty (sourcetype) typu JOB (**JOB | Windows Scheduled Task Job a JOB | Windows Scheduled Task Trigger**) reprezentujú naplánované úlohy v operačnom systéme Windows. Poskytujú informácie o automatizovaných procesoch, ktoré sa vykonávajú na základe časových alebo systémových triggerov. Ich význam spočíva najmä v detekcii mechanizmov perzistencie útočníkov. V datasetoch sa tieto artefakty objavujú len v prípadoch, kde boli plánované úlohy aktívne využívané. To znamená, že ich výskyt je silne závislý od konkrétneho scenára. Napriek tomu môžu obsahovať veľmi hodnotné informácie o škodlivých aktivitách. Rozlišujeme pritom samotné definície úloh (Job) a ich spúšťacie mechanizmy (Trigger). Ich kombinovaná analýza umožňuje rekonštruovať časovanie útokov. Nevýhodou je ich nekonzistentná prítomnosť naprieč datasetmi. Preto sú vhodné najmä pre špecializované analýzy.

LOG artefakty

Logy predstavujú jednu z najdôležitejších kategórií forenzných dát. Zahŕňajú systémové logy, logy inštalácie zariadení (**LOG | Setup API Log**) a monitorovanie využitia zdrojov. Tieto artefakty (sourcetype) poskytujú chronologický záznam udalostí v systéme. Ich výhodou je vysoká informačná hodnota a relatívne častý výskyt v datasetoch. Na druhej strane sa ich štruktúra a obsah líšia v závislosti od konfigurácie systému. Niektoré logy, ako napríklad **LOG | System Resource Usage Monitor**, sú dostupné len v určitých verziách operačného systému

Windows. To spôsobuje ich nekonzistentný výskyt. Logy môžu byť tiež objemné a obsahovať šum. Ich efektívne využitie preto vyžaduje predspracovanie a filtráciu. Napriek týmto obmedzeniam sú kľúčové pre rekonštrukciu incidentov. V kontexte tejto analýzy sú niektoré logy zaradené medzi hraničné kvôli ich variabilite.

PE (Portable Executable) artefakty

Artefakty (sourcetype) typu **PE | PE/COFF file** poskytujú informácie o spustiteľných súboroch v systéme. Obsahujú údaje ako čas kompilácie, importované knižnice či štruktúru binárneho súboru. Tieto informácie sú dôležité pri analýze škodlivého softvéru. Ich výskyt v datasetoch je však podmienený prítomnosťou exe súborov. V niektorých scenároch sa preto nevyskytujú vôbec. Napriek tomu majú vysokú hodnotu pri reverznom inžinierstve a atribúcii (stotožnení) útokov. Umožňujú identifikovať podozrivé alebo neštandardné binárne súbory. Nevýhodou je ich obmedzená univerzálnosť. Nie všetky datasetové scenáre obsahujú relevantné binárne artefakty. Preto sú vhodné skôr pre špecializované analýzy. V globálnych modeloch majú doplnkovú úlohu.

REG artefakty

Artefakty (sourcetype) uložené v rámci registra operačného systému Windows predstavujú rozsiahlu a veľmi dôležitú kategóriu forenzných dát v prostredí operačného systému Windows. Obsahujú informácie o konfigurácii systému, používateľskej aktivite a spúšťaní aplikácií. Artefakty (sourcetype) ako **REG | AppCompatCache Registry Key** alebo **REG | Background Activity Moderator Registry Entry** poskytujú cenné údaje o behu programov. Špecifické kľúče, napríklad **REG | Microsoft Outlook MRU Registry Key**, sú však viazané na konkrétne aplikácie. To spôsobuje ich nízky výskyt v datasetoch.

Artefakty (sourcetype) uložené v rámci registra operačného systému Windows sú veľmi detailné, ale zároveň fragmentované. Ich analýza vyžaduje znalosti štruktúry tohto registra. Výhodou je ich schopnosť zachytiť historické správanie systému. Nevýhodou je ich závislosť od konkrétneho operačného systému a jeho konfigurácie. Preto nie všetky kľúče sú dostupné vo všetkých datasetoch. V tejto analýze sú preto niektoré artefakty (sourcetype) nachádzajúce sa v registry operačného systému Windows klasifikované ako hraničné.

WEBHIST artefakty

Webové artefakty reprezentujú aktivitu používateľa v internetových prehliadačoch. V tomto prípade ide najmä o artefakty Internet Explorer (**WEBHIST | MSIE Cache File URL record**). Obsahujú informácie o navštívených URL, cache súboroch a reláciách. Ich význam spočíva v rekonštrukcii používateľského správania. V moderných datasetoch sa však tieto artefakty vyskytujú zriedkavo. Dôvodom je pokles používania Internet Exploreru. Preto sú dostupné len v obmedzenom počte datasetov. Napriek tomu môžu byť v niektorých prípadoch veľmi hodnotné. Ich analýza môže odhaliť komunikáciu s podozrivými doménami. Nevýhodou je ich nízka univerzálnosť. V kontexte tejto štúdie sú preto klasifikované ako hraničné.

4.2 Návrh atribútov pomocou LLM

Druhú fázu výberu vhodných atribútov do maticovej reprezentácie predstavuje návrh atribútovo pomocou veľkých jazykových modelov (LLM). Na základe predchádzajúcej analýzy sme navrhli prvotný prompt pre veľké jazykové modely, ktorého cieľom bolo automatizovane generovať návrhy binárnych atribútov reprezentujúcich jednotlivé forenzné záznamy. Prompt pre tieto jazykové modely bol formulovaný nasledovne:

Prvotný prompt:
"Načítaj súbor, ktorý predstavuje ukážku 5 datasetov z CTF pre účely digitálnej forenznej analýzy. Dataset obsahuje nasledujúce stĺpce: datetime, source, MACB, filename, desc, extra, dataset. Datetime je časová pečiatka, source je zdroj údajov, MACB predstavujú časové pečiatky pre zdrojové údaje, dataset je označenie konkrétneho datasetu a polia desc a extra obsahujú vlastné údaje. Ide o zdroj údajov WinEVTX. Na základe týchto údajov a znalostí o výstupe z nástroja Plaso *pre tento zdroj údajov vygeneruj návrh binárnych atribútov, ktoré by reprezentovali tento typ záznamov. Pre každý binárny atribút uveď názov, krátky popis atribútu a spôsob extrakcie z jednotlivých polí v jazyku Python (pre Jupyter Notebook)."*

Na základe tohto prvotného promptu boli následne vytvorené špecializované prompty pre jednotlivé typy zdrojov (*sourcetype*), pričom ich štruktúra bola jednotná a reflektovala špecifiká konkrétneho dátového zdroja. Ako príklad uvádzame prompt pre zdroj *WinEVTX* (Windows Event Logs):

"**WinEVTX**
Máš k dispozícii dataset výstupov z nástroja PLASO pre zdroj údajov WinEVTX (Windows Event Logs). Každý riadok predstavuje jeden forenzný záznam obsahujúci nasledovné polia:

- datetime: časová pečiatka udalosti,
- source: typ zdroja (napr. EVT),
- MACB: časové označenia operácií (Modified, Accessed, Changed, Birth),
- filename: názov súboru zdroja udalosti,
- desc: opis udalosti (vrátane textu z Windows Event Logu),
- extra: doplnkové atribúty ako message_identifier, recovered, sha256,
- dataset: identifikátor konkrétneho CTF datasetu.

Navrhni **sadu binárnych atribútov** vhodných na reprezentáciu týchto záznamov pre účely strojového učenia alebo forenznej klasifikácie. Pre každý atribút uveď:

1. **Názov atribútu** (v štýle názvov premenných),
2. **Stručný popis** toho, čo atribút znamená,
3. **Spôsob extrakcie** v jazyku Python vhodný pre Jupyter Notebook.

Zameraj sa na atribúty, ktoré by mohli:

- charakterizovať typ udalosti,
- identifikovať dôležité alebo podozrivé správanie,

- súvisieť s anomáliami alebo obnovovanými záznamami,
- indikovať konkrétne logy (napr. systémové, sieťové),
- reflektovať typické štruktúry v dátach PLASO/EVTX.”

Analogickým spôsobom boli pripravené prompty aj pre ďalšie typy zdrojov dát (*sourcetype-y)*:

- EVT – WinWVTX,
- FILE - File Entry Shell Item,
- FILE - File Stat,
- FILE – NTFS File Stat,
- FILE – NTFS USN Change,
- LNK – Windows Shortcut,
- LOG – WinPrefetch,
- OLECF – OLECF Dest List Entry,
- OLECF – OLECF Item,
- WEBHIST – Chrome Cache,
- WEBHIST – Chrome Cookies,
- WEBHIST – Chrome History,
- WEBHIST – MSIE WC Container rec,
- WEBHIST – MSIE WC Containers re,
- WEBHIST – MSIE WebCache cookies,
- REG – BagMRU Registry Key,
- REG – MRUList Registry Key,
- REG – MRUListEx Registry Key,
- REG_Registry Key – BagMRU,
- REG_Registry Key – MRUList,
- REG_Registry Key – MRUListEx,
- REG_Registry Key – Network Driv
- REG_Registry Key – Run Key,
- REG_Registry Key – Service,
- REG_Registry Key – Typed URLs,
- REG_Registry Key – USB Entries,
- REG_Registry Key – User Account,
- REG_Registry Key – UserAssist,
- REG_Registry Key – Winlogon,
- REG_Registry Key Shutdown Entry,
- REG_Registry Key,
- REG_Run_Run – Once Registry Key,
- REG_Service_Driver Configuration,
- REG_Shutdown Registry Key,
- REG_Task Cache Registry Key,
- REG_Task Cache,
- REG_Typer URLs Registry Key,
- REG_USB Registry Key,

- REG_USBStor Registry Key.
- REG_User Account Information Re,
- REG_UserAssist Registry Key,
- REG_Winlogon Registry Key.

Takto definované prompty boli následne použité na generovanie návrhov atribútov pomocou viacerých veľkých jazykových modelov, konkrétne: ChatGPT 4o[1], Gemini 2.5[2] a Grok 2[3]. Cieľom bolo porovnať výstupy jednotlivých modelov a identifikovať spoločné vzory v navrhovaných atribútoch.

Výsledkom tohto kroku bola množina kandidátnych binárnych atribútov pre každý analyzovaný *sourcetype*, pričom ich počet sa líšil v závislosti od typu dát a použitého modelu. Súhrn počtov atribútov pre jednotlivé kombinácie zdrojov a modelov je uvedený v Tabuľke 2.

Sourcetype	Počet všetkých atribútov	Počet atribútov vybraných pomocou Chat/GPT 4o	Počet atribútov vybraných pomocou Gemini	Počet atribútov vybraných pomocou Grok
EVT - WinEVTX	**25**	**10**	**12**	**11**
FILE - File Entry Shell Item	**21**	**10**	**13**	**15**
FILE - File Stat	**20**	**14**	**14**	**15**
FILE - NTFS File Stat	**25**	**13**	**15**	**18**
FILE - NTFS USN Change	**24**	**10**	**12**	**14**
LNK - Windows Shortcut	**23**	**7**	**12**	**12**
LOG - WinPrefetch	**14**	**4**	**5**	**10**
OLECF - OLECF Dest List Entry	**13**	**10**	**7**	**8**
OLECF - OLECF Item	**16**	**10**	**11**	**12**
WEBHIST - Chrome Cache	**15**	**10**	**11**	**13**
WEBHIST - Chrome Cookies	**21**	**10**	**11**	**14**
WEBHIST - Chrome History	**19**	**10**	**11**	**11**
WEBHIST - MSIE WC Container rec	**20**	**9**	**7**	**9**

[1] Model ChatGPT 4o, informácie o modeli dostupné na webovej stránke: https://developers.openai.com/api/docs/models/gpt-4o
[2] Model Gemini 2.5, informácie o modeli dostupné na webovej stránke: https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash
[3] Model Grok 2, informácie o modeli dostupné na webovej stránke: https://x.ai/news/grok-2

WEBHIST - MSIE WC Containers re	**20**	**8**	**10**	**15**
WEBHIST - MSIE WebCache cookies	**20**	**8**	**9**	**16**
REG - BagMRU Registry Key	**18**	**10**	**7**	**10**
REG - MRUList Registry Key	**21**	**10**	**6**	**10**
REG_MRUListEx Registry Key	**10**	**7**	**5**	**8**
REG_Registry Key – BagMRU	**13**	**6**	**6**	**10**
REG_Registry Key – MRUList	**15**	**8**	**6**	**10**
REG_Registry Key – MRUListEx	**12**	**5**	**4**	**8**
REG_Registry Key – Network Driv	**15**	**6**	**4**	**8**
REG_Registry Key – Run Key	**16**	**9**	**8**	**11**
REG_Registry Key – Service	**21**	**10**	**5**	**10**
REG_Registry Key – Typed URLs	**10**	**5**	**5**	**10**
REG_Registry Key - USB Entries	**12**	**8**	**9**	**10**
REG_Registry Key - User Account	**11**	**6**	**5**	**10**
REG_Registry Key - UserAssist	**19**	**7**	**7**	**10**
REG_Registry Key - Winlogon	**18**	**5**	**6**	**8**
REG_Registry Key Shutdown Entry	**16**	**6**	**6**	**10**
REG_Registry Key	**16**	**5**	**10**	**8**
REG_Run_Run Once Registry Key	**21**	**10**	**11**	**10**
REG_Service_Driver Configuratio	**21**	**9**	**9**	**13**
REG_Shutdown Registry Key	**19**	**5**	**7**	**10**
REG_Task Cache Registry Key	**21**	**6**	**8**	**10**
REG_Task Cache	**11**	**6**	**6**	**10**
REG_Typed URLs Registry Key	**12**	**7**	**7**	**8**

REG_USB Registry Key	**17**	**9**	**9**	**10**
REG_USBStor Registry Key	**14**	**5**	**6**	**8**
REG_User Account Information Re	**15**	**7**	**6**	**8**
REG_UserAssist Registry Key	**13**	**7**	**7**	**10**
REG_Winlogon Registry Key	**20**	**8**	**10**	**10**
SPOLU	**769**	**355**	**370**	**477**

Table 2 - Súhrn počtu atribútov pre jednotlivé kombinácie zdrojov

4.3 Konsolidácia a výber atribútov

V tretej a poslednej fáze výberu vhodných atribútov pre maticovú reprezentáciu sme pristúpili k implementácii extrakcie navrhnutých atribútov. Pôvodný zámer spočíval v tom, že do finálneho spracovania budú zahrnuté iba tie atribúty, ktoré boli identifikované aspoň dvoma nezávislými jazykovými modelmi. Tento prístup mal zabezpečiť vyššiu mieru robustnosti a eliminovať menej relevantné alebo náhodne generované atribúty.

Napriek tejto redukcii sa však ukázalo, že výsledný počet atribútov zostáva vysoký, čo viedlo k zvýšenej výpočtovej náročnosti a komplikáciám pri ďalšom spracovaní dát. Okrem toho sa ukázalo, že manuálna implementácia extrakčných pravidiel pre veľké množstvo atribútov je časovo náročná a náchylná na chyby.

Z uvedených dôvodov sme pristúpili k ďalšiemu kroku, ktorým bola redukcia dimenzionality dát. Na tento účel bola zvolená metóda *Principal Component Analysis (PCA)*, ktorá umožňuje transformovať pôvodný priestor atribútov do nižšie-dimenzionálneho priestoru pri zachovaní čo najväčšieho množstva variability v dátach. Tento prístup umožnil efektívnejšie spracovanie dát a zároveň eliminoval redundanciu medzi jednotlivými atribútmi.

5 Redukcia dimenzie dát

5.1 Úvod do redukcie dimenzie dát

V oblasti digitálnej forenznej analýzy sa často pracuje s vysokodimenzionálnymi dátami, napríklad so záznamami systémových udalostí, logmi alebo forenznými artefaktmi extrahovanými z rôznych zdrojov. Tieto dáta môžu obsahovať desiatky až stovky atribútov, pričom ich počet ďalej rastie pri agregácii dát do formátu supertimeline. Takto získané dáta môžu viesť k zvýšenej výpočtovej náročnosti, preučeniu modelov (overfitting) a zhoršenej interpretovateľnosti výsledkov. Preto sa využívajú metódy redukcie dimenzie, ktorých cieľom je znížiť počet atribútov pri zachovaní čo najväčšieho množstva relevantnej informácie [4].

Redukciu dimenzie môžeme vo všeobecnosti rozdeliť na dve hlavné skupiny:

- supervised (riadené) metódy a
- unsupervised (neriadené) metódy.

5.1.1 Supervised metódy redukcie dimenzie

Supervised metódy využívajú informáciu o cieľovej premennej a snažia sa vybrať alebo transformovať atribúty tak, aby čo najlepšie prispievali k predikcii.

Medzi najčastejšie používané prístupy patria:

- **Rozhodovacie stromy** - prirodzene vykonávajú výber atribútov počas procesu učenia, pričom menej relevantné atribúty sú eliminované.
- **Logistická regresia s reguláciou** - pri použití L1 regulácie dochádza k sparsifikácii modelu, čo vedie k eliminácii menej významných atribútov [3].
- **Linear Discriminant Analysis (LDA)** - projektuje dáta do nižšieho priestoru tak, aby maximalizovala separáciu medzi triedami.

V kontexte digitálnej forenznej analýzy sa tieto prístupy využívajú napríklad pri klasifikácii udalostí alebo identifikácii škodlivých aktivít, avšak ich použitie je podmienené dostupnosťou anotovaných dát [3].

5.1.2 Unsupervised metódy redukcie dimenzie

Na rozdiel od supervised prístupov, unsupervised metódy nevyužívajú informáciu o triedach. Ich cieľom je odhaliť štruktúru dát alebo odstrániť redundanciu medzi atribútmi.

Medzi hlavné prístupy patria:

- **Clustering (zhlukovanie)** - metódy ako k-means umožňujú identifikovať skupiny podobných udalostí alebo správania v dátach.

- **Metódy pre vizualizáciu dát (napr. t-SNE)** - používajú sa najmä na vizualizáciu vysokodimenzionálnych forenzných dát.
- **Principal Component Analysis (PCA)** - patrí medzi najpoužívanejšie metódy redukcie dimenzie. Transformuje pôvodné atribúty do nového priestoru hlavných komponentov, ktoré zachytávajú maximálnu variabilitu dát [5,6].

5.2 Principal Component Analysis (PCA)

Metóda **Principal Component Analysis (PCA)** patrí medzi najvýznamnejšie techniky redukcie dimenzie v rámci unsupervised prístupov. Jej hlavnou výhodou je schopnosť transformovať pôvodné atribúty do menšieho počtu hlavných komponentov, ktoré zachytávajú väčšinu variability dát bez potreby využitia cieľovej premennej [5,6]. PCA zároveň prispieva k eliminácii multikolinearity medzi atribútmi, keďže výsledné komponenty sú navzájom ortogonálne, a teda nekorelované.

V rámci tejto analýzy nie je k dispozícii cieľová premenná (target), ktorá by umožňovala aplikáciu supervised metód redukcie dimenzie. Z tohto dôvodu je potrebné využiť unsupervised prístupy, ktoré nevyžadujú označené dáta. PCA je v tomto kontexte vhodnou voľbou, keďže umožňuje efektívne odstrániť redundanciu medzi atribútmi a zároveň zachovať podstatnú informáciu obsiahnutú v dátach [6].

Z tohto dôvodu bola aplikovaná redukcia dimenzie pomocou PCA, a to samostatne pre jednotlivé artefakty, resp. sourcetypes. Tento prístup umožňuje lepšie zachytiť špecifické charakteristiky jednotlivých typov dát, keďže rôzne forenzné artefakty môžu obsahovať odlišné distribúcie atribútov a vzory správania. V prípade, že by bola PCA aplikovaná na všetky artefakty spoločne, mohlo by dôjsť k skresleniu výsledných komponentov, keďže dominantné vzory z jedného typu dát by mohli potlačiť významné charakteristiky iných artefaktov. Takýto prístup by mohol viesť k strate relevantnej informácie a zníženiu kvality následnej analýzy.

Pri aplikácii PCA na analyzovaný dataset došlo k redukcii počtu atribútov z pôvodných 769 na 218 hlavných komponentov, pričom bola zachovaná podstatná časť variability dát. Táto redukcia prispieva k zníženiu výpočtovej náročnosti následných analýz a zároveň k zlepšeniu prehľadnosti dát.

Na základe uvedených vlastností bola metóda PCA zvolená ako hlavný nástroj redukcie dimenzie v tejto práci.

5.2.1 Návrh a implementácia PCA modelu

Proces redukcie dimenzie pomocou metódy Principal Component Analysis (PCA) pozostával z viacerých na seba nadväzujúcich krokov, ktorých cieľom bolo transformovať pôvodný vysokodimenzionálny dataset do nižšieho priestoru pri zachovaní čo najväčšieho množstva informácie.

V prvom kroku boli vstupné dáta podrobené štandardizácii. Tento krok je nevyhnutný, keďže PCA je citlivá na mierku jednotlivých atribútov. Bez štandardizácie by atribúty s väčším rozsahom hodnôt mali neúmerne veľký vplyv na výsledné komponenty. Dáta boli preto transformované tak, aby mali nulovú strednú hodnotu a jednotkovú smerodajnú odchýlku.

Následne bol na takto upravené dáta aplikovaný PCA model bez obmedzenia počtu komponentov. Cieľom tohto kroku bolo analyzovať rozloženie vysvetlenej variability medzi jednotlivými hlavnými komponentmi a získať prehľad o tom, koľko komponentov je potrebných na zachovanie podstatnej časti informácie obsiahnutej v dátach.

Na základe výsledkov bol vytvorený tzv. scree plot, ktorý zobrazuje podiel vysvetlenej variability pre jednotlivé komponenty. Tento graf bol využitý na určenie optimálneho počtu komponentov, pričom ako kritérium bola zvolená hranica 95 % kumulatívnej vysvetlenej variability. Na základe tejto analýzy bol stanovený konečný počet komponentov použitých v ďalšom spracovaní.

Po určení optimálneho počtu komponentov bol PCA model opätovne aplikovaný, tentokrát s obmedzením na zvolený počet hlavných komponentov. Výsledkom tejto transformácie bol nový priestor príznakov, v ktorom jednotlivé komponenty predstavujú lineárne kombinácie pôvodných atribútov.

V ďalšom kroku boli analyzované váhy (tzv. loadings), ktoré vyjadrujú príspevok jednotlivých pôvodných atribútov k jednotlivým hlavným komponentom. Na základe absolútnych hodnôt týchto váh bol pre každý komponent zostavený rebríček atribútov podľa ich významnosti. Tento rebríček umožňuje identifikovať, ktoré atribúty najviac prispievajú k variabilite zachytenej konkrétnym komponentom.

Takto navrhnutý postup umožňuje nielen efektívnu redukciu dimenzie dát, ale aj zachovanie interpretovateľnosti modelu prostredníctvom analýzy príspevkov pôvodných atribútov k jednotlivým komponentom.

5.2.2 Výsledky PCA

V tejto kapitole sú prezentované výsledky redukcie dimenzie pomocou Principal Component Analysis (PCA) aplikovanej na dáta jednotlivých forenzných artefaktov zo supertimeline. PCA bola vykonaná samostatne pre každý artefakt, aby bolo možné zachytiť špecifické charakteristiky a variabilitu jednotlivých typov dát. Pre každý artefakt bol určený optimálny počet hlavných komponentov na základe scree plotu a hranice kumulatívnej vysvetlenej variability 90-95 %. Následne boli analyzované váhy pôvodných atribútov (loadings) a zostavené rebríčky najdôležitejších atribútov pre každý artefakt, ktoré poskytujú prehľad o tom, ktoré atribúty najviac prispievajú k variabilite dát. Výsledky sú prezentované v tabuľkách (Tabuľka 3 – Tabuľka 7) samostatne pre jednotlivé zdroje údajov (CTF), čo umožňuje porovnať význam atribútov naprieč rôznymi dátovými sadami.

Artefakt	Pôvodné atribúty	Počet komponentov	% vysvetlenej variability	Počet vybraných atribútov
WinEVTX	27	15	93,7%	17
File_entry_shell_item	23	5	92,2%	8
NTFS_file_stat	27	11	93,1%	15
File_stat	22	11	95,9%	15
Windows_Shortcut	25	4	93,6%	7
Amcache_Registry_Entry				
AppCompatCache_Registry_Key				
BagMRU_Registry_Key	20	3	93,7%	5
Chrome_Cache	17	8	94,8%	10
Chrome_Cookies	22	7	93,8%	8
Chrome_History	21	4	94,8%	5
MRUList_Registry_Key	23	1	100%	1
MRUListEx_Registry_Key	12	4	100%	4
MSIE_WebCache_container_record	22	7	97,2%	10
MSIE_WebCache_containers_record	23	5	95,3%	8
MSIE_Cache_File_URL_record				
Network_Connection_Registry_Key				
OLECF_Dest_list_entry	15	6	97,4%	7
OLECF_Item	17	1	100%	1
OLECF_Summary_Info	2	1	100%	1
Open_XML_Metadata				
PE_Compilation_time				
PE_COFF_file	2	1	100%	1
Registry_Key	16	11	93,7%	12
Registry_Key___BagMRU				
Registry_Key___UserAssist				
Registry_Key___Run_Key				
Registry_Key___MRUList				
Registry_Key___MRUListEx				
Registry_Key___Typed_URLs				
Run_Run_Once_Registry_Key	22	3	99,9%	8
System	2	1	100%	1
Service_Driver_Configuration_Registry_Key				
Setup_API_Log				

Shutdown_Registry_Key				
Task_Cache_Registry_Key	20	5	91,7%	6
USB_Registry_Key	16	1	100%	1
User_Account_Information_Registry_Key				
WinPrefetch	16	8	97,6%	10
Winlogon_Registry_Key				

Table 2 - PCA pre prípad Magnet_CTF_2019_Windows_Desktop

Artefakt	Pôvodné atribúty	Počet komponentov	% vysvetlenej variability	Počet vybraných atribútov
WinEVTX	27	15	91,7%	17
File_entry_shell_item	23	5	94,4%	7
NTFS_file_stat	27	11	95,4%	15
File_stat	22	11	96,9%	15
Windows_Shortcut	25	5	95,5%	9
Amcache_Registry_Entry				
AppCompatCache_Registry_Key	2	1	100%	1
BagMRU_Registry_Key	20	4	92,9%	6
Chrome_Cache	17	8	93,6%	10
Chrome_Cookies	22	9	90,6%	11
Chrome_History	21	5	89,4%	6
MRUList_Registry_Key	23	3	100%	4
MRUListEx_Registry_Key	12	2	100%	2
MSIE_WebCache_container_record	22	8	98,1%	10
MSIE_WebCache_containers_record	23	4	94,0%	5
MSIE_Cache_File_URL_record	2	1	100%	1
Network_Connection_Registry_Key	2	1	100%	1
OLECF_Dest_list_entry	15	3	92,0%	4
OLECF_Item	17	3	100%	3
OLECF_Summary_Info	2	1	100%	1
Open_XML_Metadata	2	1	100%	1
PE_Compilation_time				
PE_COFF_file	2	1	100%	1
Registry_Key	16	13	97,2%	14
Registry_Key___BagMRU				

Registry_Key___UserAssist				
Registry_Key___Run_Key				
Registry_Key___MRUList				
Registry_Key___MRUListEx				
Registry_Key___Typed_URLs				
Run_Run_Once_Registry_Key	22	4	92,7%	8
System	2	1	100%	1
Service_Driver_Configuration_Registry_Key	2	1	100%	1
Setup_API_Log	2	1	100%	1
Shutdown_Registry_Key	8	1	100%	1
Task_Cache_Registry_Key	20	4	100%	4
USB_Registry_Key	16	2	100%	2
User_Account_Information_Registry_Key	2	1	100%	1
WinPrefetch	16	8	95,6%	11
Winlogon_Registry_Key	2	1	100%	1

Table 3 - PCA pre prípad Magnet_CTF_2020_Windows desktop

Artefakt	Pôvodné atribúty	Počet komponentov	% vysvetlenej variability	Počet vybraných atribútov
WinEVTX	27	16	95,9%	17
File_entry_shell_item	23	5	95,1%	8
NTFS_file_stat	27	10	95,5%	14
File_stat	22	11	92,7%	16
Windows_Shortcut	25	5	97,1%	8
Amcache_Registry_Entry	2	1	100%	1
AppCompatCache_Registry_Key				
BagMRU_Registry_Key	20	3	100%	4
Chrome_Cache	17	8	94,7%	9
Chrome_Cookies	22	10	97,8%	11
Chrome_History	21	7	98,3%	8
MRUList_Registry_Key	23	2	100%	5
MRUListEx_Registry_Key	12	1	100%	1
MSIE_WebCache_container_record	22	5	94,5%	6
MSIE_WebCache_containers_record	23	3	97,2%	5
MSIE_Cache_File_URL_record				

Network_Connection_Registry_Key				
OLECF_Dest_list_entry	15	5	95,9%	7
OLECF_Item	17	1	100%	1
OLECF_Summary_Info				
Open_XML_Metadata				
PE_Compilation_time				
PE_COFF_file	2	1	100%	1
Registry_Key	16	11	97,8%	12
Registry_Key___BagMRU				
Registry_Key___UserAssist				
Registry_Key___Run_Key				
Registry_Key___MRUList				
Registry_Key___MRUListEx				
Registry_Key___Typed_URLs				
Run_Run_Once_Registry_Key	22	3	98,3%	10
System	2	1	100%	1
Service_Driver_Configuration_Registry_Key				
Setup_API_Log				
Shutdown_Registry_Key				
Task_Cache_Registry_Key	20	5	94,4%	6
USB_Registry_Key	16	1	100%	1
User_Account_Information_Registry_Key				
WinPrefetch	16	9	97,3%	12
Winlogon_Registry_Key				

Table 4 - PCA pre prípad Magnet_CTF_2022_Windows laptop

lArtefakt	Pôvodné atribúty	Počet komponentov	% vysvetlenej variability	Počet vybraných atribútov
WinEVTX	27	18	95,0%	22
File_entry_shell_item	23	3	94,7%	6
NTFS_file_stat	27	10	95,9%	15
File_stat	22	10	94,4%	15
Windows_Shortcut	25	3	96,7%	7
Amcache_Registry_Entry				
AppCompatCache_Registry_Key				

BagMRU_Registry_Key				
Chrome_Cache				
Chrome_Cookies				
Chrome_History				
MRUList_Registry_Key				
MRUListEx_Registry_Key				
MSIE_WebCache_container_record	22	4	98,7%	9
MSIE_WebCache_containers_record	23	4	96,3%	6
MSIE_Cache_File_URL_record				
Network_Connection_Registry_Key				
OLECF_Dest_list_entry	15	3	100%	3
OLECF_Item	17	1	100%	1
OLECF_Summary_Info				
Open_XML_Metadata				
PE_Compilation_time				
PE_COFF_file				
Registry_Key	16	12	97,7%	13
Registry_Key___BagMRU	12	3	95,0%	4
Registry_Key___UserAssist	18	3	100%	4
Registry_Key___Run_Key	2	1	100%	1
Registry_Key___MRUList				
Registry_Key___MRUListEx				
Registry_Key___Typed_URLs				
Run_Run_Once_Registry_Key				
System	2	1	100%	1
Service_Driver_Configuration_Registry_Key				
Setup_API_Log				
Shutdown_Registry_Key				
Task_Cache_Registry_Key				
USB_Registry_Key				
User_Account_Information_Registry_Key				
WinPrefetch				
Winlogon_Registry_Key				

Table 5 - PCA pre prípad SSS DC

Artefakt	Pôvodné atribúty	Počet komponentov	% vysvetlenej variability	Počet vybraných atribútov
WinEVTX	27	18	93,5%	20
File_entry_shell_item	23	4	93,0%	5
NTFS_file_stat	27	11	95,8%	15
File_stat	22	11	94,5%	16
Windows_Shortcut	25	4	100%	6
Amcache_Registry_Entry				
AppCompatCache_Registry_Key				
BagMRU_Registry_Key				
Chrome_Cache				
Chrome_Cookies				
Chrome_History				
MRUList_Registry_Key				
MRUListEx_Registry_Key				
MSIE_WebCache_container_record	22	1	100%	1
MSIE_WebCache_containers_record				
MSIE_Cache_File_URL_record				
Network_Connection_Registry_Key				
OLECF_Dest_list_entry	15	4	95,1%	5
OLECF_Item	17	1	100%	1
OLECF_Summary_Info				
Open_XML_Metadata				
PE_Compilation_time	2	1	100%	1
PE_COFF_file				
Registry_Key	16	12	95,2%	13
Registry_Key___BagMRU	12	3	97,2%	4
Registry_Key___UserAssist	18	2	100%	2
Registry_Key___Run_Key	2	1	100%	1
Registry_Key___MRUList	2	1	100%	1
Registry_Key___MRUListEx	2	1	100%	1
Registry_Key___Typed_URLs	2	1	100%	1
Run_Run_Once_Registry_Key				
System	2	1	100%	1
Service_Driver_Configuration_Registry_Key				
Setup_API_Log				

Shutdown_Registry_Key				
Task_Cache_Registry_Key				
USB_Registry_Key				
User_Account_Information_ Registry_Key				
WinPrefetch	16	8	96,5%	12
Winlogon_Registry_Key				

Table 6 - PCA pre prípad SSS Desktop

Výsledky sú prezentované v piatich samostatných tabuľkách (Tabuľka 3 – Tabuľka 7), pričom každá tabuľka zodpovedá konkrétnemu typu zdroja (sourcetype), na ktorom bola PCA aplikovaná nezávisle. Tento prístup umožňuje detailnejšie zachytiť špecifiká jednotlivých typov dát, keďže rôzne sourcetypy obsahujú odlišné množiny artefaktov a atribútov.

V niektorých prípadoch sa v tabuľkách nachádzajú nevyplnené hodnoty. Tie môžu nastať z dvoch dôvodov:

- buď daný artefakt nebol prítomný v konkrétnom sourcetype, alebo
- síce prítomný bol, avšak nebolo možné naň aplikovať PCA. Táto situácia nastáva najmä v prípadoch, keď atribúty daného artefaktu vykazujú nulovú variabilitu (t. j. obsahujú konštantné hodnoty), a preto neprispievajú k vysvetleniu variability dát a nie je možné z nich extrahovať hlavné komponenty.

Táto neúplnosť teda nepredstavuje chybu spracovania, ale je dôsledkom charakteru analyzovaných forenzných dát.

Z výsledkov uvedených v tabuľkách vyplýva, že aplikácia PCA viedla k výraznej redukcii počtu atribútov pri zachovaní vysokej miery vysvetlenej variability vo všetkých analyzovaných artefaktoch. Vo väčšine prípadov bolo možné reprezentovať pôvodné dáta pomocou výrazne menšieho počtu hlavných komponentov, pričom si model zachoval viac ako 90 % informácie obsiahnutej v dátach.

Miera redukcie sa však medzi jednotlivými artefaktmi líši, čo naznačuje rozdielnu mieru redundancie a komplexnosti ich atribútov. Artefakty s vyšším stupňom redukcie obsahujú pravdepodobne viac korelovaných alebo redundantných premenných, zatiaľ čo artefakty s nižšou mierou redukcie si vyžadujú väčší počet komponentov na zachytenie variability, čo poukazuje na ich vyššiu informačnú rôznorodosť.

Analýza váh atribútov (loadings) zároveň umožnila identifikovať kľúčové atribúty, ktoré najviac prispievajú k variabilite dát. Tieto atribúty môžu byť považované za najvýznamnejšie z pohľadu ďalšieho spracovania, napríklad pri klasifikácii alebo detekcii anomálií.

Celkovo možno konštatovať, že PCA predstavuje efektívny nástroj na redukciu dimenzie forenzných dát zo supertimeline, pričom umožňuje zachovať podstatnú časť informácie a zároveň zjednodušiť následné analytické úlohy.

5.3 Spojenie dát z PCA

Z dôvodu obmedzených výpočtových kapacít nebolo možné spracovať všetkých 769 atribútov naraz. Preto bol prístup založený na aplikácii metódy Principal Component Analysis (PCA) samostatne pre každý *sourcetype*.

Cieľom však nebolo pracovať s viacerými oddelenými tabuľkami, ale vytvoriť jednotnú reprezentáciu dát. Po výbere relevantných komponentov pomocou PCA bol preto na jednotlivé datasety aplikovaný preprocessing, ktorého výsledkom bola extrakcia 178 vybraných atribútov pre každý dataset.

Následne boli tieto atribúty zlúčené do jednej spoločnej tabuľky. Pre zachovanie informácie o pôvode jednotlivých záznamov boli zároveň pridané pomocné atribúty typu *mask_sourcetype*, ktoré explicitne označujú príslušnosť záznamu ku konkrétnemu zdroju dát.

V Tabuľke 8 je uvedený prehľad počtu vybraných atribútov pre jednotlivé artefakty (sourcetype) po aplikácii PCA, pričom tabuľka zároveň ilustruje mieru redukcie dimenzie v rámci jednotlivých zdrojov údajov.

Artefakt	Pôvodné atribúty	Počet vybraných atribútov
WinEVTX	27	23
File_entry_shell_item	23	9
NTFS_file_stat	27	15
File_stat	22	16
Windows_Shortcut	25	10
Amcache_Registry_Entry	2	1
AppCompatCache_Registry_Key	2	1
BagMRU_Registry_Key	20	6
Chrome_Cache	17	10
Chrome_Cookies	22	11
Chrome_History	21	9
MRUList_Registry_Key	23	9
MRUListEx_Registry_Key	12	5
MSIE_WebCache_container_record	22	13
MSIE_WebCache_containers_record	23	9
MSIE_Cache_File_URL_record	2	1

Network_Connection_Registry_Key	2	1
OLECF_Dest_list_entry	15	8
OLECF_Item	17	3
OLECF_Summary_Info	2	1
Open_XML_Metadata	2	1
PE_Compilation_time	2	1
PE_COFF_file	2	1
Registry_Key	16	14
Registry_Key___BagMRU	12	4
Registry_Key___UserAssist	18	5
Registry_Key___Run_Key	2	1
Registry_Key___MRUList	2	1
Registry_Key___MRUListEx	2	1
Registry_Key___Typed_URLs	2	1
Run_Run_Once_Registry_Key	22	10
System	2	1
Service_Driver_Configuration_Registry_Key	2	1
Setup_API_Log	2	1
Shutdown_Registry_Key	8	1
Task_Cache_Registry_Key	20	7
USB_Registry_Key	16	2
User_Account_Information_Registry_Key	2	1
WinPrefetch	16	12
Winlogon_Registry_Key	2	1

Table 7 - Zjednotenie po súboroch (image)

Tabuľka 9 obsahuje finálny zoznam všetkých atribútov po spojení dát, pričom výsledná maticová reprezentácia pozostáva zo 178 atribútov. Táto tabuľka zahŕňa kompletný prehľad atribútov použitých v ďalšom spracovaní a predstavuje jednotnú reprezentáciu dát naprieč všetkými dátovými sadami.

stĺpec	Artefakt	Popis atribútu
access_count_gt_1	WEBHIST - MSIE WebCache Container record	Objekt bol navštívený alebo použitý viac než raz.
accessed_appdata	REG - MRUList Registry Key	Prístup k súborom v AppData (často zaujímavé z pohľadu malware).
accessed_local_file	WEBHIST - MSIE WebCache Container record	Záznam reprezentuje prístup k lokálnemu súboru cez file:// cestu.

accessed_network_path	REG - MRUList Registry Key	Prístup k súboru na sieťovej ceste (UNC path).
accessed_sensitive_docs	REG - MRUList Registry Key	Indikuje prístup k potenciálne citlivým dokumentom (Office súbory, PDF alebo výskyt slova „password“).
accessed_sensitive_file	WEBHIST - MSIE WebCache Container record	Záznam obsahuje prístup k potenciálne citlivému súboru alebo obsahu.
account_creation	WinEVTX	Explicitne identifikuje vytvorenie nového používateľského účtu.
account_deletion	WinEVTX	Explicitne identifikuje odstránenie používateľského účtu.
container_id	WEBHIST - MSIE WC Containers record	Unikátny identifikátor WebCache kontajnera.
contains_appdata_path	REG - Run_Run Once Registry Key	Spúšťaný súbor sa nachádza v AppData.
contains_cab	FILE - File Entry Shell Item	CAB archív (balíčky, aktualizácie).
contains_cur	FILE - File Entry Shell Item	Cursor súbor (.cur), skôr zriedkavé.
contains_inf	FILE - File Entry Shell Item	INF súbor (inštalačné skripty, často pre ovládače).
contains_maintenance_task	REG - Task Cache Registry Key	Úloha súvisí s údržbou systému.
contains_path_downloads	REG - MRUListEx Registry Key	Záznam obsahuje cestu do priečinka Downloads.
contains_path_recent	REG - MRUListEx Registry Key	Súbor pochádza z priečinka Recent (nedávno otvorené položky).
contains_ppd	FILE - File Entry Shell Item	Printer Description file.
contains_secret_path	WEBHIST - MSIE WebCache Container record	Indikuje lokálnu cestu file://, ktorá obsahuje slovo secret.
contains_sensitive_string	REG - MRUListEx Registry Key	Záznam obsahuje citlivé kľúčové slová (napr. password, credentials, secret).
contains_update_task	REG - Task Cache Registry Key	Úloha súvisí s aktualizáciami.
contains_vbs	FILE - File Entry Shell Item	VBScript súbor (často zneužívaný malwareom).
ContainsCmdCommand	REG - Registry Key	Registry obsahuje CMD príkaz.

ContainsPowershellCommand	REG - Registry Key	Registry obsahuje PowerShell príkaz.
ContainsScriptExtension	REG - Registry Key	Registry odkazuje na skript (BAT, PS1, VBS, JS).
event_powershell	WinEVTX	PowerShell aktivita.
event_registry	WinEVTX	Operácie nad Windows Registry.
has_exe_target	LNK - Windows Shortcut	Shortcut smeruje na spustiteľný súbor .exe.
has_hostname_reference	OLECF - Dest List Entry	Záznam obsahuje referenciu na hostname (môže indikovať sieťový zdroj).
has_ip_url	WEBHIST - MSIE WebCache Container record	URL je zapísaná priamo pomocou IP adresy namiesto doménového mena.
has_macb_accessed	FILE - File Entry Shell Item	Súbor bol otvorený (Accessed).
has_macb_modified	FILE - File Entry Shell Item	Súbor bol modifikovaný.
has_modified_flag	FILE - File Entry Shell Item	Súbor bol modifikovaný (Modified timestamp).
	FILE - File Stat	
has_modified_time	LNK - Windows Shortcut	Prvý znak MACB je M, teda záznam indikuje modifikáciu v modified timestamp.
has_persistence_trigger	REG - Task Cache Registry Key	Úloha má periodický alebo podmienený trigger (napr. denne, pri nečinnosti).
has_ps1_target	LNK - Windows Shortcut	Shortcut smeruje na PowerShell skript .ps1.
has_sha256	FILE - File Stat	Súbor má dostupný SHA256 hash (dôležité pre identifikáciu a threat intel).
has_transition_type	WEBHIST - Chrome History	Záznam obsahuje informáciu o type prechodu (napr. klik, redirect, typed URL).
has_unknown_shell_item	REG - Registry Key - BagMRU	Indikuje, že záznam obsahuje neznámy alebo nerozpoznaný shell item typ.
HasAandBNoMC	LNK - Windows Shortcut	Shortcut má access aj birth timestamp, ale nemá modified ani changed timestamp.
HasCommandArguments	REG - Run_Run Once Registry Key	Spustiteľný súbor obsahuje argumenty (napr. program.exe -arg).

is_accessed	FILE - File Stat	Súbor bol otvorený (access).
	WEBHIST - Chrome Cookies	Cookie bola použitá (accessed), teda stránka ju aktívne využila.
is_ad_service	WEBHIST - Chrome Cache	Zdroj je spojený s reklamnými službami (Google Ads, DoubleClick).
is_admin_activity	WinEVTX	Aktivita spojená s administrátorskými účtami.
is_advertising_cookie	WEBHIST - Chrome Cookies	Cookie používaná na reklamné účely (ad targeting).
is_allocated	FILE - File Entry Shell Item	Súbor je alokovaný (existuje v súborovom systéme).
is_analytics_cookie	WEBHIST - Chrome Cookies	Cookie používaná na analytiku (napr. Google Analytics).
is_apis_google	WEBHIST - Chrome Cache	Zdroj využíva Google API endpointy.
is_application_experience	WinEVTX	Udalosti služby Application Experience.
is_bing_domain	WEBHIST - Chrome Cookies	Cookie pochádza z domény Bing.
is_browser_exe	LOG - WinPrefetch	Záznam patrí webovému prehliadaču.
is_cdn_usage	WEBHIST - Chrome Cache	Zdroj je načítaný cez Content Delivery Network (napr. Akamai, CDN servery).
is_cmd	REG - Registry Key - UserAssist	Spustenie príkazového riadku (Command Prompt).
is_cmd_or_script	FILE - File Entry Shell Item	Indikuje spustenie príkazového riadku alebo skriptov (CMD, BAT, PowerShell).
is_content_cache	WEBHIST - MSIE WC Containers record	Kontajner obsahuje cache obsah (INetCache – uložené webové zdroje).
is_content_container	WEBHIST - MSIE WC Containers record	Kontajner typu Content (webový obsah – HTML, obrázky, skripty).
is_desktop_or_downloads	REG - Registry Key - BagMRU	Záznam odkazuje na priečinok Desktop alebo Downloads.
is_error_event	WinEVTX	Chybové alebo neúspešné udalosti.
is_event_log_related	WinEVTX	Udalosti súvisiace s Windows Event Log službou.

is_evtx_started_state	WinEVTX	Služba prešla do bežiaceho stavu.
is_evtx_stopped_state	WinEVTX	Služba bola zastavená.
is_executable	FILE - File Stat	Spustiteľný súbor alebo systémový driver.
	REG - MRUList Registry Key	Indikuje otvorenie alebo spustenie spustiteľného súboru alebo skriptu.
	OLECF - Dest List Entry	Záznam reprezentuje spustiteľný súbor alebo skript.
is_failed_login	WinEVTX	Indikuje neúspešné pokusy o prihlásenie (napr. Event ID 4625 alebo failed logon).
is_file	FILE - File Stat	Záznam reprezentuje súbor.
is_file_entry	FILE - File Stat	Alternatívna detekcia file entry (presný formát logu).
is_filename_attr	FILE - File Entry Shell Item	NTFS atribút $FILE_NAME
is_fileshare_access	WEBHIST - MSIE WebCache Container record	Indikuje prístup k zdieľanému súboru alebo sieťovej ceste; zachytáva aj textové súbory.
is_frequent_execution	LOG - WinPrefetch	Aplikácia bola spustená veľmi často, aspoň 50-krát.
is_from_driverstore	FILE - File Entry Shell Item	Súbor sa nachádza v DriverStore (ovládače systému).
is_from_ie	WEBHIST - Chrome History	Záznam bol importovaný z Internet Exploreru.
is_from_pinned_location	FILE - File Entry Shell Item	Súbor bol spustený z pripnutého umiestnenia (Taskbar, Quick Launch).
is_from_suspicious_domain	WEBHIST - Chrome Cookies	Cookie pochádza z potenciálne rizikovej alebo spamovej domény (heuristika podľa kľúčových slov).
is_from_system32	FILE - File Entry Shell Item	Súbor pochádza zo systémového priečinka System32.
is_from_user_profile	OLECF - Dest List Entry	Súbor sa nachádza v používateľskom profile (AppData, Roaming, Users).
is_from_users_dir	FILE - File Entry Shell Item	Súbor je v používateľskom adresári.
is_google_domain	WEBHIST - Chrome Cache	Zdroj pochádza z domény google.com.

	WEBHIST - Chrome Cookies	Cookie pochádza z domény Google.
is_gstatic_domain	WEBHIST - Chrome Cache	Zdroj pochádza z domény gstatic.com (statické súbory Google služieb).
is_high_access_count	WEBHIST - MSIE WebCache Container record	Objekt má vysoký počet prístupov, viac než 10.
is_history_container	WEBHIST - MSIE WC Containers record	Kontajner pre históriu prehliadania (navštívené stránky).
is_https	WEBHIST - Chrome History	URL používa zabezpečený protokol HTTPS.
is_iecompat_container	WEBHIST - MSIE WC Containers record	Kontajner súvisiaci s kompatibilitou Internet Exploreru.
is_iecompatua_container	WEBHIST - MSIE WC Containers record	Kontajner pre user-agent kompatibilitu (IE režimy).
is_image_file	WEBHIST - Chrome Cache	Cache obsahuje obrázkový súbor.
is_in_appdata	FILE - File Stat	Súbor sa nachádza v priečinku AppData (častý cieľ malware-u).
is_in_system32	FILE - File Stat	Súbor je v systémovom priečinku System32.
is_in_temp_or_spool	FILE - File Stat	Súbor je v dočasnom alebo spool priečinku (často krátkodobé alebo podozrivé súbory).
is_in_winsxs	FILE - File Stat	Súbor je vo WinSxS (Windows Side-by-Side, legitímne systémové knižnice).
is_language_code	OLECF - OLECF Item	Položka má názov reprezentovaný 4-ciferným kódom, čo často zodpovedá jazykovým alebo identifikačným kódom v OLECF štruktúre.
is_large_file	FILE - File Stat	Súbor väčší ako ~10 MB.
is_link_file	FILE - File Entry Shell Item	Indikuje Windows shortcut (.lnk), často používaný na sledovanie spúšťania programov.
is_log_recovered	WinEVTX	Log bol obnovený po chybe alebo páde.

is_login_event	WinEVTX	Identifikuje úspešné prihlásenia používateľa (napr. Event ID 4624 alebo logon).
is_modified_after_access	WEBHIST - MSIE WebCache Container record	MACB záznam začína M, čo indikuje modifikáciu po alebo pri prístupe.
is_multiple_volumes	LOG - WinPrefetch	Spustenie je asociované s viacerými volume-mi, čo môže naznačovať prístup k externým alebo viacerým úložiskám.
is_network_event	TENTO ATRIBUT SA NEPARSUJE	
is_network_path	FILE - File Entry Shell Item	Súbor pochádza zo sieťovej cesty (napr. UNC path).
	REG - MRUListEx Registry Key	Záznam reprezentuje prístup k súboru na sieťovej ceste (UNC path).
	REG - Run_Run Once Registry Key	Spúšťanie z sieťovej alebo vzdialenej cesty.
is_network_related	WinEVTX	Sieťová konektivita a udalosti.
	REG - Registry Key - UserAssist	Indikuje, že spustený program súvisí so sieťovou komunikáciou.
is_network_share	REG - BagMRU Registry Key	Indikuje prístup na sieťové zdieľanie cez UNC cestu, napr. \\server\share.
is_night_access	WEBHIST - Chrome History	Aktivita prebehla v noci (22:00 – 06:00).
is_night_execution	LOG - WinPrefetch	Indikuje udalosti súvisiace so zmenou bezpečnostných politík systému (napr. Event ID 4719 alebo výskyt slova policy).
is_ntfs	FILE - File Stat	Súbor je na NTFS súborovom systéme.
is_ntfs_event	WinEVTX	Udalosti NTFS súborového systému.
is_obfuscated_path	LOG - WinPrefetch	Cesta obsahuje znaky obfuskácie, napr. hexadecimálne názvy adresárov alebo podozrivé sekvencie ...
is_pinned	OLECF - Dest List Entry	Súbor je pripnutý (pinned) v Jump List.

is_policy_change	WinEVTX	Indikuje udalosti súvisiace so zmenou bezpečnostných politík systému (napr. Event ID 4719 alebo výskyt slova policy).
is_powershell	REG - Registry Key - UserAssist	Spustenie PowerShellu.
is_privilege_change	WinEVTX	Označuje udalosti, kde došlo k zmene alebo použitiu privilegovaných práv (napr. Event ID 4672, 4673 alebo výskyt privilege).
is_rare_app	LOG - WinPrefetch	Aplikácia bola spustená len zriedkavo, menej než 5-krát.
is_repeated_record	WEBHIST - MSIE WC Containers record	Identifikuje kontajnery, ktoré sa vyskytujú viackrát (viac záznamov).
is_root_entry	OLECF - OLECF Item	Identifikuje koreňovú položku (Root Entry) OLECF súboru.
is_run_often	LOG - WinPrefetch	Aplikácia bola spustená opakovane, viac než 5-krát.
is_runonce_key	REG - Run_Run Once Registry Key	Kľúč typu RunOnce – vykoná sa len raz pri štarte.
is_script_file	FILE - File Stat	Skript (.bat alebo PowerShell .ps1).
	WEBHIST - Chrome Cache	JavaScript súbor.
is_secure	WEBHIST - Chrome Cookies	Cookie má nastavený atribút Secure (prenáša sa len cez HTTPS).
is_sensitive_file	OLECF - Dest List Entry	Záznam obsahuje indikáciu citlivého súboru (napr. „SECRET“, „password“).
is_service_event	WinEVTX	Udalosti súvisiace so správou služieb systému (Service Control Manager).
is_social_media	WEBHIST - Chrome History	URL patrí medzi sociálne siete.
is_specific_container	WEBHIST - MSIE WebCache Container record	Záznam patrí do špecifického WebCache kontajnera, napr. History alebo Cookies.
is_standard_info	FILE - File Entry Shell Item	NTFS atribút $STANDARD_INFORMATION.
is_suspicious_app	LOG - WinPrefetch	Indikuje spustenie nástroja často zneužívaného pri

		útokoch alebo post-exploitation.
is_suspicious_domain	WEBHIST - Chrome Cache	Doména nepatrí medzi známe/whitelistované (Google, YouTube, Facebook, Twitter), teda potenciálne menej dôveryhodná.
	WEBHIST - MSIE WebCache Container record	URL smeruje na doménu s koncovkou .ru alebo .cn, čo je heuristicky označené ako potenciálne podozrivé.
is_suspicious_extension	REG - Run_Run Once Registry Key	Spúšťaný súbor má podozrivú príponu (skripty alebo executable).
is_suspicious_keyword	WEBHIST - MSIE WebCache Container record	Záznam obsahuje podozrivé alebo citlivé kľúčové slová.
is_suspicious_task_name	REG - Task Cache Registry Key	Názov úlohy zodpovedá generickým alebo maskovacím názvom.
is_suspicious_tool	REG - Registry Key - UserAssist	Spustený program patrí medzi známe nástroje zneužívané pri útokoch (LOLbins).
is_system_hive	REG - Run_Run Once Registry Key	Kľúč patrí systémovému registry hive.
is_system_location	REG - BagMRU Registry Key	Označuje, že záznam smeruje na systémovú alebo špeciálnu shell lokalitu, napr. Control Panel, Recycle Bin alebo Network.
is_system_path	LNK - Windows Shortcut	Shortcut sa nachádza v systémovej alebo spoločnej lokalite, napr. Windows, System32 alebo ProgramData.
is_system_process	LOG - WinPrefetch	Indikuje, že prefetchnutý program patrí medzi bežné systémové procesy Windows.
is_system_util	LOG - WinPrefetch	Záznam patrí systémovému utilitnému nástroju, ktorý môže byť legitímny aj zneužitý.
is_temp_launch	LOG - WinPrefetch	Program bol spustený z dočasného priečinka TEMP.
is_terminal_services	WinEVTX	Vzdialený prístup (RDP, Terminal Services).

is_third_party	WEBHIST - Chrome Cookies	Cookie pochádza z third-party domény (nie priamo z navštíveného webu).
is_tracking_cookie	WEBHIST - Chrome Cookies	"Všeobecný indikátor tracking cookies (kombinácia analytiky a reklamy). "
is_txt_file	OLECF - Dest List Entry	Súbor je textový (.txt).
is_unallocated	FILE - File Entry Shell Item	"Súbor bol zmazaný (nealokovaný priestor). "
is_unpinned	OLECF - Dest List Entry	Súbor nie je pripnutý (bežný recent item).
is_unusual_location	FILE - File Entry Shell Item	Identifikuje súbory v neobvyklých (často zneužívaných) lokalitách ako AppData alebo Temp, mimo štandardných systémových priečinkov.
is_url_visit	WEBHIST - MSIE WebCache Container record	Indikuje, že záznam obsahuje návštevu webovej URL adresy cez HTTP alebo HTTPS.
is_usb_hub	REG - USB Registry Key	Indikuje, že zariadenie je USB hub (rozbočovač).
is_user_account_event	WinEVTX	Označuje udalosti týkajúce sa používateľských účtov (napr. vytvorenie účtu – Event ID 4720).
is_user_location	REG - MRUList Registry Key	Súbor sa nachádza v používateľských priečinkoch (Documents, Desktop, Downloads, AppData).
is_user_profile_path	REG - Registry Key - BagMRU	Záznam obsahuje cestu do používateľského profilu v tvare C:\Users\<username>.
is_user_profile_shortcut	LNK - Windows Shortcut	Shortcut sa nachádza v používateľskom profile, napr. v Users, AppData alebo Roaming.
is_xml_file	FILE - File Stat	XML súbor (často konfigurácie).
is_youtube_domain	WEBHIST - Chrome Cookies	Cookie pochádza z domény YouTube.
IsAdministratorProfile	WEBHIST - MSIE WC Containers record	Kontajner patrí administrátorskému účtu.
IsCSS	WEBHIST - Chrome Cache	CSS súbor (štýly web stránky).

IsDynamicLibrary	FILE - File Stat	Dynamická knižnica (.dll).
IsInAppData	FILE - File Entry Shell Item	Súbor sa nachádza v priečinku AppData.
IsInRoamingAppData	FILE - File Entry Shell Item	"Súbor sa nachádza v lokálnom AppData (nie roaming). "
IsSuspiciousPath	REG - Registry Key	Cesta obsahuje podozrivé prvky (temp, náhodné názvy, GUID).
IsSystemHive	REG - Registry Key	Kľúč patrí systémovému hive (HKLM).
IsUserHive	REG - Registry Key	Kľúč patrí používateľskému hive (HKCU).
key_autorun_related	REG - Registry Key	Indikuje autorun/autostart mechanizmy.
key_contains_run	REG - Registry Key	Registry kľúč obsahuje Run/RunOnce – klasický mechanizmus persistence.
key_contains_winlogon	REG - Registry Key	Kľúč súvisí s Winlogon – veľmi silný persistence bod.
key_new_registry_branch	REG - Registry Key	Indikuje vytvorenie alebo prístup k novému vetveniu registry.
key_suspicious_value	REG - Registry Key	Registry obsahuje podozrivé spustiteľné nástroje.
macb_anomaly	LNK - Windows Shortcut	Špecifický neštandardný MACB vzor .A.B, ktorý môže indikovať anomálne časové správanie shortcutu.
macb_is_only_access_birth	LNK - Windows Shortcut	MACB obsahuje iba access a/alebo birth príznaky bez modifikácie či zmeny metadata.
MissingMInMACB	LNK - Windows Shortcut	Shortcut nemá v MACB zázname príznak modifikácie (M).
ModifiesImagePath	REG - Registry Key	Indikuje zmenu cesty k spustiteľnému súboru (často služby alebo driver).
ModifiesStartType	REG - Registry Key	Zmena typu spúšťania služby (napr. auto-start).
multiple_entries	REG - MRUList Registry Key	Záznam obsahuje viac MRU položiek (viacero naposledy použitých objektov).

opened_control_panel	REG - BagMRU Registry Key	Indikuje, že používateľ otvoril alebo prehliadal položku Control Panel.
opened_recycle_bin	REG - BagMRU Registry Key	Indikuje, že používateľ otvoril alebo prehliadal Recycle Bin.
OpenedExcelFile	REG - MRUList Registry Key	Indikuje otvorenie Excel súboru (.xls, .xlsx).
OpenedWordDocument	REG - MRUList Registry Key	Indikuje otvorenie Word dokumentu (.doc, .docx).
potential_malware	WinEVTX	Potenciálne škodlivé správanie.
RunsRundll32	REG - Run_Run Once Registry Key	Spúšťa sa rundll32.exe – často zneužívaný nástroj na spúšťanie DLL.
suspicious_executable	REG - Run_Run Once Registry Key	Názov spustiteľného súboru napodobňuje legitímne systémové procesy.
task_starts_at_boot	REG - Task Cache Registry Key	Úloha sa spúšťa pri štarte systému.
task_starts_at_logon	REG - Task Cache Registry Key	Úloha sa spúšťa pri prihlásení používateľa.
unknown_shell_item_type	REG - BagMRU Registry Key	Záznam obsahuje neznámy alebo neidentifikovaný shell item typ.
UsesQuotesInPath	REG - Run_Run Once Registry Key	Cesta k súboru je v úvodzovkách.
visit_to_search_engine	WEBHIST - Chrome History	Návšteva stránky výsledkov vyhľadávania (Google alebo Bing).
visited_microsoft	WEBHIST - Chrome History	Návšteva domény Microsoft.
visited_msn	WEBHIST - Chrome History	Návšteva domény MSN.

Tabuľka 8 - Popis atribútov vybraných PCA

6 Zhrnutie

Po realizácii procesu spojenia dát obsahuje výsledná dátová reprezentácia **178 atribútov**, rozšírených o **40 maskovacích stĺpcov** reprezentujúcich jednotlivé sourcetype-y. Tieto stĺpce umožňujú explicitne zachytiť zdrojový kontext každého záznamu a podporujú jeho ďalšiu interpretáciu.

Vzhľadom na nasledujúci krok spracovania, konkrétne agregácie, bolo potrebné detailne analyzovať obsah novovzniknutých datasetov. Na tento účel bola vytvorená sumarizačná tabuľka, ktorá pre každý atribút a každý dataset (CTF) obsahuje počet výskytov hodnôt 0, 1 a NaN. Táto tabuľka (summary_table) je súčasťou prílohy k tomuto výstupu - „D15 - Model na extrakciu digitálnych stôp do maticovej reprezentácie – výsledky“ a nachádza sa v zošite „summary_table“.

Tabuľka zahŕňa všetky spracované datasety a bola vytvorená jednotným spôsobom pre každý z nich, čo umožňuje ich vzájomné porovnanie. Poskytuje tak komplexný prehľad o distribúcii jednotlivých atribútov, ich prítomnosti (hodnota 1), absencii (hodnota 0) a chýbajúcich hodnôt naprieč všetkými datasetmi. Zároveň slúži ako podklad pre ďalšie spracovanie, najmä pri návrhu agregačných prístupov a práci s neúplnými dátami.

Súčasťou tabuľky je zároveň aj slovný popis jednotlivých atribútov a ich interpretácia z pohľadu digitálnej forenznej analýzy, vrátane zhodnotenia ich relevancie pre vyšetrovanie. Je však dôležité zdôrazniť, že nízka individuálna relevancia atribútu neznamená jeho bezvýznamnosť. V kombinácii s inými atribútmi môže poskytovať významnú kontextovú informáciu a prispievať k identifikácii komplexnejších vzorcov správania.

Tabuľka „summary_table“ tak slúži nielen ako kvantitatívny prehľad dát, ale aj ako interpretačný podklad pre ďalšie spracovanie, najmä pri návrhu agregačných prístupov a práci s neúplnými dátami.

Výskyt NaN hodnôt je dôsledkom predchádzajúceho spracovania dát – konkrétne situácií, keď daný atribút prislúcha určitému sourcetype, ktorý však nie je v konkrétnom riadku zastúpený. Inými slovami, tieto chýbajúce hodnoty neindikujú absenciu informácie, ale nerelevantnosť atribútu pre daný záznam. Ide teda o štrukturálne chýbajúce hodnoty, ktoré vznikajú prirodzene pri spájaní heterogénnych dátových zdrojov do jednotnej reprezentácie.

Tento typ NaN hodnôt je potrebné interpretovať odlišne od klasických chýbajúcich dát spôsobených napríklad nedostupnosťou alebo stratou informácie. V našom prípade NaN explicitne signalizuje, že daný atribút nie je pre konkrétny artefakt definovaný, a preto by nemal byť pri ďalšom spracovaní (napr. agregácii alebo modelovaní) považovaný za nulovú hodnotu.

Správna interpretácia týchto hodnôt je kľúčová najmä pri návrhu agregačných funkcií, kde je potrebné rozhodnúť, či sa NaN bude ignorovať, alebo bude mať špecifický význam v rámci výpočtu. Tento aspekt priamo ovplyvňuje výslednú reprezentáciu dát a môže mať významný dopad na následnú analýzu a interpretáciu výsledkov.

7 Bibliografia

[1] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn, and L. Chang, "A novel anomaly detection scheme based on principal component classifier," in Proc. IEEE Foundations and New Directions of Data Mining Workshop, 2003.

[2] M. Ring, S. Wunderlich, D. Grüdl, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," Computers & Security, vol. 86, pp. 147–167, 2019.

[3] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," Journal of Computational Science, vol. 25, pp. 152–160, 2018.

[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Computing Surveys, vol. 41, no. 3, pp. 1–58, 2009.

[5] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and Intelligent Laboratory Systems, vol. 2, no. 1–3, pp. 37–52, 1987.

[6] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," Philosophical Transactions of the Royal Society A, vol. 374, no. 2065, 2016.

[7] Plaso (log2timeline), 2022, [online] Available: https://github.com/log2timeline/plaso.

8 Prílohy

Tabuľka ADFIR-D15-KPB3-01-table predstavuje prílohu k analýze dát a obsahuje dva samostatné zošity, ktoré poskytujú prehľad o štruktúre zdrojových dát a vlastnostiach extrahovaných atribútov:

- **sourcetype_source_counts_summar** – sumarizačná tabuľka zachytávajúca počet výskytov jednotlivých typov zdrojov (sourcetype) naprieč analyzovanými datasetmi (CTF), ktorá slúži na identifikáciu dominantných, hraničných a menej vhodných zdrojov údajov pre ďalšie spracovanie.
- **summary_table** – sumarizačná tabuľka binárnych atribútov po predspracovaní, ktorá pre každý atribút a dataset obsahuje počty hodnôt 0, 1 a NaN, vrátane popisu atribútov a ich interpretácie z pohľadu digitálnej forenznej analýzy.