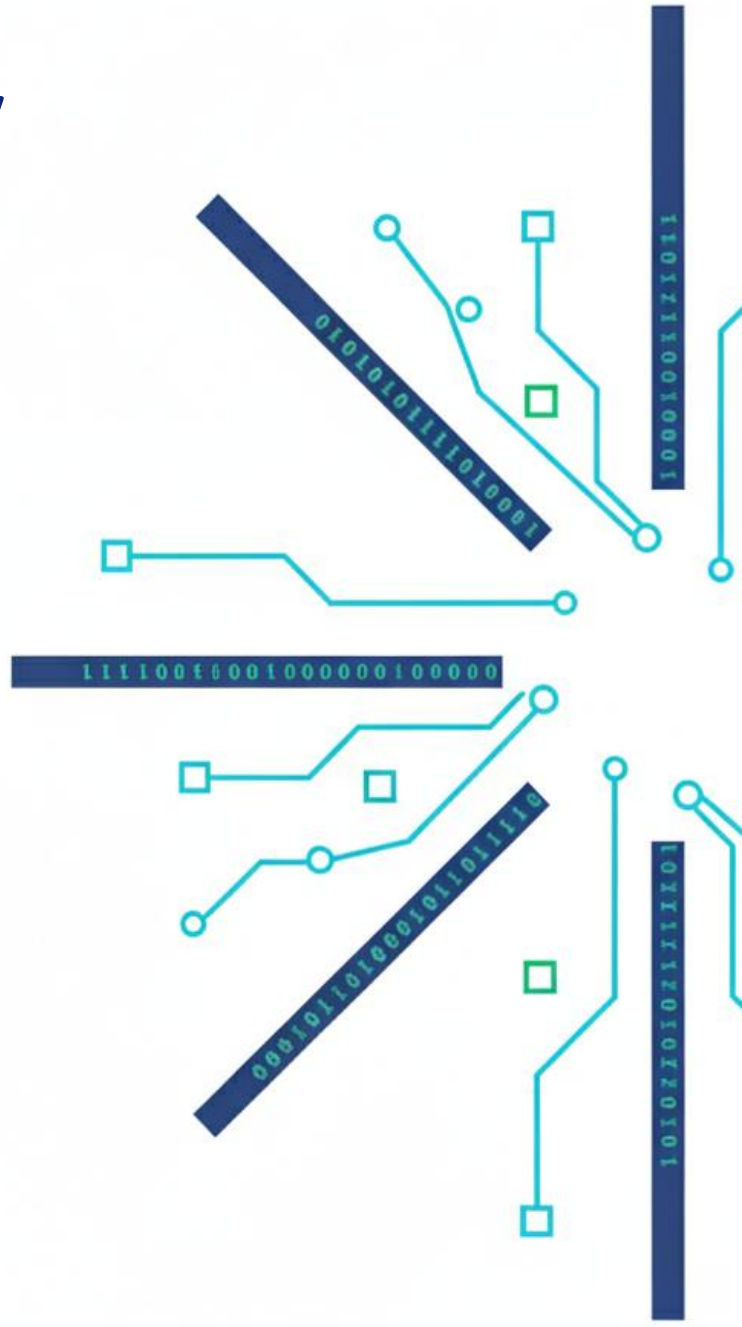
D14 – Syntetický dataset



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.

Osnova

1 Popis projektu

Projekt **Automatizácia digitálnej forenzie a reakcie na incidenty** (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 rieši jednu z kľúčových výziev v oblasti kybernetickej bezpečnosti a informačnej bezpečnosti – ako spracovať obrovské množstvo digitálnych stôp, 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:

- **Urýchlené riešenie incidentov v oblasti kybernetickej bezpečnosti**. Projekt ADFIR zavádza automatizované prístupy k zberu, spracovaniu a analýze digitálnych stôp. V dôsledku toho 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

Dôležitým aspektom výskumu digitálnych forenzných údajov je schopnosť vytvoriť súbor údajov, ktorý spĺňa určité očakávania a požiadavky. Vo všeobecnosti neexistuje žiadny súbor údajov, ktorý by bolo možné použiť na všetky výskumné účely v oblasti digitálnej forenznej analýzy [1, 2]. Pri používaní, vytváraní a zdieľaní dátových súborov čelia výskumníci viacerým výzvam. Je dôležité poznamenať, že vo všeobecnosti takéto dátové súbory v tejto oblasti chýbajú alebo chýba dokumentácia a formálny popis ich konštrukcie [3, 4]. Na účely nášho výskumu a projektu potrebujeme pracovať s dátovým súborom, ktorý opisuje reálne scenáre, ktoré sa môžu vyskytnúť v skutočných bezpečnostných incidentoch. Cieľom je vytvoriť vhodný súbor údajov na porovnanie metód analýzy digitálnych stôp, ktorý by sa dal použiť na vyšetrovanie rôznych problémov. Vo všeobecnosti existujú rôzne súbory údajov, s ktorými výskumníci pracujú, ale buď ich nezdieľajú, alebo nie všetky sú použiteľné v tomto type výskumu. Existujú aj rámce, ktoré sa dajú použiť na generovanie súborov údajov, ale nie je zaručené, že súbor údajov je správne vytvorený a spĺňa všetky podmienky potrebné pre tento typ výskumu.

Sme si vedomí, že takáto dátová sada **nemusí pokrývať všetky situácie,** ktoré sa môžu vyskytnúť v reálnom prostredí. V súlade s cieľmi projektu je prezentovaný výstup – dátová sada – **vytvorený predovšetkým z dátových sád používaných v súťažiach CTF a je primerane doplnený realistickými údajmi**. Z týchto dôvodov sa dátová sada skladá z viacerých komplementárnych častí:

- **Dataset vytvorený simuláciou rôznych techník útočníkov**, ktorý je podrobnejšie opísaný v kapitole 4.
- **Dataset vytvorený z obrazov diskov pochádzajúcich zo súťaží CTF**, ktorý je podrobnejšie opísaný v kapitole 5. Táto časť datasetu obsahuje:
    - vyhodnotené a spracované výstupy z forenzných nástrojov (modifikovaný súbor údajov EZ tools) a
    - modifikovaný **vložený súbor údajov**.

Štruktúra tohto výstupu sleduje **schému článku v časopise *Data in Brief***, ktorý sa zameriava na prezentáciu a popis výskumných dátových súborov. Stručný prehľad dátového súboru je uvedený v **tabuľke 1**.

Téma	Informatika
Špecifická oblasť	Kyberbezpečnosť, digitálna forenzná analýza.
Typ údajov	Tabuľkové údaje (súbory CVS) obsahujúce údaje, databáza SQL a vložené údaje z viacerých forenzných artefaktov operačného systému Windows.
Zber údajov	Údaje boli získané z dvoch zdrojov, kategorizovaných podľa ich typu. Údaje týkajúce sa rôznych techník útočníkov boli zhromaždené zo simulovaného prostredia poskytnutého spoločnosťou IstroSec a následne zhromaždené pomocou nástroja Athena. Druhý typ údajov pozostáva zo spracovaných obrázkov zariadení (počítačov a notebookov) stiahnutých z rôznych webových stránok a úložísk súťaží

	CTF. Účelom týchto súťaží CTF je poskytnúť praktické skúsenosti v oblasti digitálnej forenzie, reakcie na incidenty a vyhľadávania hrozieb. Konkrétne sme použili údaje z prípadu s názvom *The Stolen Szechuan Sauce* z portálu DFIR Madness, údaje z Magnet CTF (2019, 2022) a údaje z prípadu úniku údajov NIST. Tieto súbory údajov sú voľne dostupné na príslušných portáloch a široko sa používajú na školiace účely, pričom podporujú začiatočníkov aj profesionálov v oblasti reakcie na incidenty a digitálnej forenzie.
Umiestnenie zdroja údajov	Údaje týkajúce sa rôznych techník útočníkov boli uložené v laboratóriu IstroSec. Druhý typ údajov pochádza z obrazových súborov disku doménového radiča a tiež z obrazových súborov disku z pracovnej plochy z prípadu s názvom The Stolen Szechuan Sauce. Konkrétne sme použili súbory DC01 Disk Image (E01) a Desktop Disk Image (E01). Ďalšie obrazové súbory sú Magnet CTF 2019 Windows Desktop, Magnet CTF 2022 Windows Laptop a NIST Data Leakage Case.
Dostupnosť údajov	Názov úložiska: Mendeley Data – ADFIR forensic dataset Identifikačné číslo údajov: 65g9gm8zrd Priama URL adresa údajov: 10.17632/65g9gm8zrd.1
Súvisiace výskumné články	Hennelová, Z., Marková, E., & Sokol, P. (2025, august). Vplyv anti-forenzných techník na digitálnu forenznú analýzu založenú na údajoch: Prípadová štúdia detekcie anomálií. V *Medzinárodnej konferencii o dostupnosti, spoľahlivosti a bezpečnosti* (s. 131-148). Cham: Springer Nature Switzerland. Antoni, Ľ., Sokol, P., Krišáková, S. P., Kotlárová, D., Krídlo, O., & Krajči, S. (2025). Formálna analýza konceptov a atribútové závislosti prípadu úniku dát NIST.

Tab. 1 – Tabuľka špecifikácií

3 Hodnota údajov

Z hľadiska digitálnej forenznej analýzy je nevyhnutné venovať pozornosť hodnote a potenciálu dostupných údajov, pretože tvoria základ pre efektívnu analýzu bezpečnostných incidentov. Hodnota prezentovaného súboru údajov sa dá identifikovať v nasledujúcich aspektoch:

Tento súbor údajov umožňuje efektívne **uplatňovanie metód analýzy údajov a strojového učenia** v oblasti digitálnej forenznej analýzy v prostredí operačného systému Windows s použitím súborového systému NTFS. Súbor údajov je zložený z kombinácie údajov zo simulovaných útokov a údajov pochádzajúcich zo súťaží Capture The Flag (CTF), ktoré spolu poskytujú realistický pohľad na správanie systému a aktivity útočníkov. Keďže NTFS je najpoužívanejší súborový systém v operačnom systéme Windows, tieto údaje predstavujú kľúčový príklad aplikácie analytických metód a metód strojového učenia v oblasti kybernetickej bezpečnosti, konkrétne v digitálnej forenznej analýze. V rámci nášho výskumu boli na tieto údaje aplikované metódy ako formálna konceptová analýza [5, 6] a detekcia anomálnych hodnôt [7].

Dataset je určený pre výskumníkov zameraných na aplikáciu analýzy dát a strojového učenia v kybernetickej bezpečnosti, ako aj pre odborníkov, ktorí sa zaujímajú o **automatizáciu procesov reakcie na bezpečnostné incidenty** a digitálnu forenznú analýzu v operačnom systéme Windows a prostredí súborového systému NTFS.

Predstavený súbor údajov je vhodný na **vývoj automatizovaných bezpečnostných nástrojov** zameraných na identifikáciu relevantných digitálnych stôp v systémoch Windows, analýzu vzťahov medzi stopami, ich atribútmi a inými kontextovými závislosťami vyplývajúcimi z forenzných artefaktov operačného systému Windows a súborového systému NTFS.

Identifikácia digitálnych stôp a vzorov správania systémových a útočných aktivít na základe údajov zo simulovaných útočných techník, ako aj scenárov CTF v prostredí operačného systému Windows, môže významne prispieť k vývoju účinnejších nástrojov na detekciu a prevenciu kybernetických útokov. To umožňuje organizáciám účinnejšie chrániť svoje informačné systémy a údaje pred bezpečnostnými hrozbami.

Okrem retrospektívnej forenznej analýzy je súbor údajov použiteľný aj na **prediktívnu analýzu a prevenciu bezpečnostných hrozieb** v prostredí operačného systému Windows a súborového systému NTFS. Uplatnenie metód analýzy údajov a strojového učenia na kombinované údaje zo simulovaných útokov a súťaží CTF umožňuje identifikovať vzorce správania a anomálie, ktoré môžu naznačovať potenciálne bezpečnostné riziká alebo ohrozenia v rámci IT infraštruktúry.

4 Súbor údajov o technikách simulovaných útočníkov

Súbor údajov o technikách simulovaných útočníkov vytvorený v rámci tohto projektu bol navrhnutý na školenie a validáciu modelov strojového učenia (ML) zameraných na automatizáciu digitálnej forenznej analýzy. Jeho hlavným cieľom je realisticky zachytiť správanie moderných útočníkov vo **fáze po zneužití**, s osobitným dôrazom na pokročilé techniky používané skupinami APT. Na rozdiel od čisto syntetických alebo náhodne generovaných súborov údajov tento súbor údajov kombinuje automatizovanú simuláciu útokov s ručne vykonávanými pokročilými útočnými aktivitami, čím zabezpečuje vysokú úroveň realizmu a forenznej relevantnosti.

4.1 Popis údajov

Základná vrstva sieťovej a systémovej aktivity bola generovaná pomocou **platformy Cymulate**, ktorá poskytla realistické pozadie legitímnych operácií a simulovala štandardné vektory narušenia. V tomto prostredí bol vykonaný pokročilý scenár emulujúci správanie skupiny **APT-19**, ktorý pokrýval širokú škálu techník od vykonávania škodlivého kódu až po perzistenciu a obchádzanie obrany.

Súbor údajov obsahuje široké spektrum **taktík, techník a postupov (TTP)**, najmä:

- zneužitie PowerShell (stiahnutie dátovej náplne prostredníctvom Invoke-WebRequest, vykonávanie príkazov kódovaných v Base64, skryté okná PowerShell),
- vykonávanie príkazov ako služby Windows,
- process injection pomocou **Reflective DLL Injection**,
- inštalácia a spúšťanie služieb prostredníctvom PowerShell,
- modifikácie registrov (mechanizmy perzistencie, skrytie prípon súborov, manipulácia s nastaveniami týkajúcimi sa bezpečnosti),
- zneužitie legitímnych systémových binárnych súborov (**Living off the Land Binaries – LOLBINs**), vrátane použitia alternatívnych dátových tokov (ADS),
- manipulácia s premennou prostredia PATH s cieľom uniesť tok vykonávania.

Okrem tejto automatizovanej základnej línie boli ručne vložené útočné sekvencie zamerané na kľúčové fázy reťazca útokov po zneužití. **Trvalosť** bola dosiahnutá pomocou nástroja **SharPersist**, ktorý vytvoril naplánované úlohy, služby a položky registru, aby zabezpečil dlhodobý prístup cez zadné dvierka. Nasledovalo **vymenovanie prostredia a eskalácia privilégií**, vykonané pomocou **SharpUp** a **winPEAS**, ktoré generovali veľký objem systémových dotazov a výrazný „hlučný“ vzor v protokoloch, charakteristický pre agresívne prieskumy po zneužití.

Pokiaľ ide o **prístup** k **povereniam**, súbor údajov zachytáva použitie **SafetyKatz** a **SharpKatz** na extrakciu autentifikačných materiálov NTLM a Kerberos z procesu LSASS, vrátane vytvorenia a odstránenia výpisov pamäte. Fáza **obchádzania obrany** je reprezentovaná použitím **SharpKiller**, techniky obchádzania AMSI, ktorá deaktivuje skenovanie v pamäti a simuluje sofistikovaných protivníkov, ktorí sa pokúšajú zaslepiť bezpečnostné monitorovanie, než pristúpia k ďalším škodlivým akciám.

Z hľadiska **hrozbového prostredia** súbor údajov odráža súčasný posun od tradičného malvéru založeného na súboroch k technikám **Living off the Land (LotL)** a **.NET tradecraft**. Mnoho nástrojov sa načíta priamo do pamäte s minimálnym priestorom na disku, čo núti detekčné mechanizmy – a najmä modely ML – zamerať sa na behaviorálne anomálie, kontext procesu a používanie API namiesto statických podpisov súborov. Úmyselné umiestnenie nástrojov do dôveryhodných adresárov, ako je C:\Windows\Tasks, v kombinácii s legitímne vyzerajúcimi názvami súborov, ďalej podporuje výskum kontextovo orientovanej detekcie maskovacích procesov.

Vďaka zahrnutiu techník aktívne používaných štátom sponzorovanými aktérmi aj modernými skupinami **Ransomware-as-a-Service (RaaS)** má tento dataset silnú **praktickú použiteľnosť**. Umožňuje vývoj a hodnotenie detekčných prístupov založených na ML, ktoré sú schopné identifikovať bezsúborové útoky, pokročilé perzistentné hrozby a aktivity v ranom štádiu zamerané na oslabenie alebo obídenie bezpečnostných kontrol v reálnych prostrediach.

Podrobnejšie informácie o procese vytvárania súboru údajov sú uvedené v dokumente: **ADFIR – D13 – Metóda vytvárania súboru údajov.**

4.2 Schéma databázy

Tento modul definuje hierarchiu projektu, jednotlivé zbierky a identifikáciu zariadení.

Názov stĺpca	Typ údajov	Popis
Id	int	Jedinečný identifikátor projektu (PK).
Guid	uniqueidentifier	Globálne jedinečný identifikátor projektu.
DtCreatedUtc	datetime2(2)	Dátum a čas vytvorenia projektu.
DtUpdatedUtc	datetime2(2)	Dátum a čas poslednej úpravy projektu.
Name	nvarchar(128)	Názov projektu.
Code	nvarchar(250)	Interný kódový názov projektu.
ProjFolder	nvarchar(256)	Cesta k priečinku, v ktorom sú uložené údaje projektu.

Tab. 2 – Projekt

Názov stĺpca	Typ údajov	Popis
Id	int	Jedinečný identifikátor parametra (PK).
ProjectId	int	Cudzí kľúč, ktorý odkazuje na tabuľku Project.

Name	nvarchar(64)	Názov parametra.
Value	nvarchar(1024)	Hodnota parametra.

Tab. 2 – Projektový parameter

Názov stĺpca	Typ údajov	Popis
Id	int	Jedinečný identifikátor relácie (PK).
Guid	uniqueidentifier	Globálne jedinečný identifikátor relácie.
AgentVersion	bigint	Verzia agenta, ktorý vykonal zber údajov.
ProjectId	int	Cudzí kľúč, ktorý odkazuje na tabuľku Project.
HostId	int	Cudzí kľúč odkazujúci na tabuľku Host.
DtStartUtc	datetime2(2)	Čas začatia zberu údajov.
DtEndUtc	datetime2(2)	Čas ukončenia zberu údajov.
Result	int	Vrátený kód výsledku zberu (úspech/chyba).

Tab. 3 – Relácia

Názov stĺpca	Typ údajov	Popis
Id	int	Jedinečný identifikátor hosťa (PK).
Guid	uniqueidentifier	GUID zariadenia.
DtRegisteredUtc	datetime2(2)	Dátum registrácie zariadenia v systéme.
HostName	nvarchar(256)	Sieťové meno zariadenia.
HardwareId	uniqueidentifier	Jedinečné hardvérové ID (HWID).
OsId	uniqueidentifier	Identifikátor operačného systému.

Tab. 4 – Hosť

Názov stĺpca	Typ údajov	Popis
Id	int	Jedinečný identifikátor artefaktu (PK).
SessionId	int	Odkaz na reláciu Session.
ArtConfigId	int	ID konfigurácie použitej pre tento artefakt.

Deskriptor	nvarchar(512)	Pôvodný identifikátor zdroja (napr. cesta k súboru).
Checksum	bigint	Kontrolný súčet artefaktu na overenie integrity.
Status	tinyint	Stav artefaktu, ktorý sa prenáša alebo spracováva.

Tab. 5 – ArtObject (základná entita)

4.2.1 Súborový systém NTFS (SchemaMft)

Tento modul mapuje štruktúru hlavnej tabuľky súborov (MFT).

Stĺpec	Typ údajov	Popis
Id	bigint	Primárny kľúč.
ArtObjectId	int	Odkaz na ArtObject.
Signature	int	Podpis záznamu (zvyčajne „FILE“).
UpdateSeqOffset	smallint	Posun na aktualizáciu sekvenčného poľa.
Lsn	bigint	Číslo sekvencie protokolu ($LogFile).
SequenceNumber	smallint	Poradové číslo pre opätovné použitie záznamu.
ReferenceCount	smallint	Počet pevných odkazov na súbor.
Flags	smallint	Priepínače (InUse, Directory).
BaseRecordRef	bigint	Odkaz na základný záznam MFT (ak ide o rozšírenie).

Tab. 6 – MftEntry (hlavný záznam)

Stĺpec	Typ údajov	Popis
MftAttrHeaderId	Bigint	FK na hlavičke atribútu.
DtCreatedUtc	datetime2(0)	Čas vytvorenia.
DtLastModifiedUtc	datetime2(0)	Dátum úpravy obsahu (čas úpravy).
DtMftEntryLastMod...	datetime2(0)	Čas zmeny MFT (čas zmeny MFT).
DtLastAccessUtc	datetime2(0)	Dátum posledného prístupu (čas prístupu).
FileAttributeFlags	Int	Priepínače súboru (len na čítanie, skrytý, systémový...).

Tab. 7 – MftStandardInformation (časové značky a atribúty)

Stĺpec	Typ údajov	Popis
MftAttrHeaderId	bigint	FK na hlavičke atribútu.
ParentDirectoryRef	bigint	Odkaz na záznam MFT nadradeného priečinka.

DtCreatedUtc	datetime2(0)	Čas vytvorenia (z hľadiska názvu súboru).
AllocatedFileSize	bigint	Alokovaná veľkosť súboru na disku.
RealFileSize	bigint	Skutočná veľkosť dát súboru.
FileName	nvarchar(512)	Názov súboru/zložky.
Namespace	tinyint	Typ menného priestoru (POSIX, Win32, DOS).

Tab. 8 – MftFileName (názov a nadradený)

4.2.2 Výkonné artefakty: Predbežné načítanie (SchemaPrefetch)

Spracované súbory .pf používané na optimalizáciu spúšťania aplikácie.

Stĺpec	Typ údajov	Popis
Id	bigint	Primárny kľúč.
FileSize	int	Veľkosť súboru predbežného načítania.
ExecutableFilename	nvarchar(60)	Názov spustiteľného súboru (napr. CMD.EXE).
PrefetchHash	int	Hash cesty k súboru (rozlišuje rovnaké názvy od iných ciest).
RunCount	int	Celkový počet spustení aplikácie.
VolumeCount	int	Počet zväzkov, z ktorých bola aplikácia spustená.

Tab. 9 – PrefetchFileHeader

Stĺpec	Typ	Popis
PrefetchFileHeaderId	bigint	FK k hlavičke súboru predbežného načítania.
LastRunTime	datetime2(0)	Časová pečiatka jedného z posledných spustení.

Tab. 10 – PrefetchLastRunTime

Stĺpec	Typ	Popis
PrefetchStartTime	int	Čas od začiatku behu, keď bol súbor načítaný.
PrefetchDuration	int	Ako dlho trvalo načítanie súboru.
FileName	nvarchar(512)	Názov načítaného súboru (závislosti, DLL).

Tab. 11 – PrefetchFileMetrics

4.2.3 Súbory LNK a skratky (SchemaShellLink)

Podrobná dekompozícia binárnej štruktúry súborov Shell Link (LNK).

Stĺpec	Typ údajov	Popis
Id	bigint	Primárny kľúč.
LinkFlags	int	Priepínače určujúce prítomnosť iných dátových štruktúr.
FileAttributesFlags	int	Atribúty cieľového súboru (skrytý, systémový...).
CreationTime	datetime2(3)	Čas, kedy bol cieľový súbor vytvorený.
WriteTime	datetime2(3)	Čas zápisu cieľového súboru.
FileSize	int	Veľkosť cieľového súboru.

Tab. 12 – ShellLinkHeader

Názov stĺpca	Typ údajov	Povoliť nulové hodnoty	Popis
Id	bigint	Nie	Jedinečný identifikátor (PK).
ShellLinkHeaderId	bigint	Áno	Cudzí kľúč k ShellLinkHeader.
NameString	nvarchar(256)	Áno	Voliteľný popis odkazu.
RelativePath	nvarchar(256)	Áno	Relatívna cesta k cieľovému súboru.
WorkingDir	nvarchar(256)	Áno	Pracovný adresár na spustenie aplikácie.
CommandLineArguments	nvarchar(256)	Áno	Argumenty príkazového riadku (kľúč pre forenznú analýzu).
IconLocation	nvarchar(256)	Áno	Cesta k súboru s ikonou.

Tab. 13 – ShellLinkStringData (textové údaje)

Stĺpec	Typ údajov	Popis
MachineId	nvarchar(256)	NetBIOS názov počítača, na ktorom bolo vytvorené prepojenie.

Droid1	uniqueidentifier	ID zväzku.
Droid2	jedinečný identifikátor	ID súboru.

Tabuľka 14 – ShellLinkTrackerData (distribuované sledovanie odkazov)

4.2.4 Protokoly udalostí (SchemaEventLog)

Tento modul sa používa na ukladanie spracovaných protokolov udalostí systému Windows (.evtx).

Názov stĺpca	Typ údajov	Popis
Id	bigint	Jedinečný identifikátor riadku (PK).
ArtObjectId	int	Cudzí kľúč k ArtObject (zdrojový súbor).
EventId	int	ID udalosti (napr. 4688, 4624).
Verzia	tinyint	Verzia štruktúry udalosti.
Qualifiers	smallint	Kvalifikátory udalosti (často 0).
[Level]	tinyint	Úroveň závažnosti (Info, Upozornenie, Chyba).
Task	smallint	Kategória úlohy v rámci poskytovateľa.
Opcode	tinyint	Operačný kód (napr. Info, Start, Stop).
Keywords	bigint	Kľúčové slovo Bitová maska (úspech/neúspech auditu).
RecordId	bigint	Pôvodné poradové číslo záznamu v protokole.
Názov poskytovateľa	nvarchar(255)	Názov zdroja (napr. Microsoft-Windows-Security-Auditing).
ProviderId	uniqueidentifier	GUID poskytovateľa.
ChannelName	nvarchar(255)	Názov kanála (bezpečnosť, systém, aplikácia).
ProcessId	int	PID procesu, ktorý zaznamenal udalosť.
ThreadId	int	ID vlákna.
ComputerName	varchar(255)	Názov počítača, na ktorom udalosť vznikla.
UserId	varbinary(128)	SID používateľa (binárne).

DtCreatedUtc	datetime2(2)	Presný čas výskytu udalosti.
ActivityId	uniqueidentifier	ID aktivity pre koreláciu (ETW).
RelatedActivityId	uniqueidentifier	ID súvisiacej aktivity.
Description	nvarchar(MAX)	Úplný textový popis udalosti (ak bol získaný).

Tab. 15 – EvtRecord

Názov stĺpca	Typ údajov	Popis
Id	bigint	Jedinečný identifikátor vlastnosti (PK).
EvtRecordId	bigint	Cudzí kľúč v EvtRecord.
Name	nvarchar(128)	Názov parametra (napr. TargetImage, IpAddress).
Value	nvarchar(MAX)	Hodnota parametra (napr. C:\Windows\System32\cmd.exe).

Tab. 16 – EvtRecordProperty (dynamické údaje)

4.2.5 Asistent kompatibility programov (SchemaPca)

Stĺpec	Typ údajov	Popis
ExecutablePath	nvarchar(512)	Úplná cesta k spustenej aplikácii.
LastExecutedUtc	datetime2(3)	Čas posledného spustenia.

Tab. 17 – PcaLaunchDictionary

Stĺpec	Typ údajov	Popis
RuntimeUtc	datetime2(3)	Čas spustenia/záznamu.
ExecutablePath	nvarchar(512)	Cesta k súboru.
ExitCode	nvarchar(64)	Návratový kód na konci procesu.
FileVersion	varchar(32)	Verzia súboru.

Tab. 18 – PcaGeneralDatabase

4.2.6 Kôš (SchemaRecycleBin)

Stĺpec	Typ údajov	Popis
DeletedFileSize	bigint	Veľkosť súboru v bajtoch.
DtDeletedUtc	datetime2(3)	Čas, kedy bol súbor odstránený (presunutý do koša).
DeletedFileName	nvarchar(512)	Pôvodný názov a cesta k odstránenému súboru.

Tab. 19 – RecycleBinFileEntry

4.2.7 Hlásenie chýb systému Windows (SchemaWer)

Modul analýzy chybových správ systému Windows. Hlavná tabuľka je WerReport.

Názov stĺpca	Typ údajov	Popis
Id	int	Primárny kľúč správy.
ArtObjectId	int	Prepojenie s tabuľkou ArtObject (zdrojový súbor).
Verzia	int	Verzia správy WER.
Typ udalosti	varchar(32)	Typ udalosti (napr. „APPCRASH“, „BEX“).
EventTimeUtc	datetime2(3)	Čas výskytu udalosti v UTC.
ReportType	int	Číselný typ správy.
UploadTimeUtc	datetime2(3)	Čas, kedy bola správa odoslaná na server Microsoft.
ReportIdentifier	uniqueidentifier	Jedinečný identifikátor GUID identifikujúci túto správu.
OriginalFilename	nvarchar(512)	Pôvodný názov súboru, ktorý spôsobil chybu.
AppPath	nvarchar(512)	Cesta k aplikácii, ktorá sa zrútila.
AppName	nvarchar(512)	Názov aplikácie.

Tab. 20 – WerReport

Názov stĺpca	Typ údajov	Popis
Id	int	Identifikátor záznamu.
Report Id	int	Odkaz na hlavnú správu (WerReport).
[Key]	nvarchar(512)	Názov kľúča metadát (napr. „ProcessId“).

Value	nvarchar(512)	Hodnota (napr. číslo procesu).

Tab. 21 – WerReportProcessMetadata

Názov stĺpca	Typ údajov	Popis
Id	int	Identifikátor záznamu.
Report Id	int	Odkaz na hlavnú správu.
Path	nvarchar(512)	Úplná cesta k načítanému modulu/knižnici.

Tab. 22 – WerReportLoadedModule

Názov stĺpca	Typ	Popis
Id	int	Identifikátor záznamu.
Report Id	int	Odkaz na hlavnú správu.
Name	nvarchar(512)	Názov parametra (napr. „AppVersion“).
Value	nvarchar(512)	Hodnota parametra (napr. „1.0.0.0“)

Tab. 23 – WerReportSig

Názov stĺpca	Typ údajov	Popis
Id	int	ID záznamu.
Report Id	int	Odkaz na WerReport.
Name	nvarchar(512)	Názov.
Value	nvarchar(512)	Hodnota.

Tabuľka 24 – WerReportDynamicSiq

Názov stĺpca	Typ údajov	Popis
Id	int	ID záznamu.
Report Id	int	Odkaz na WerReport.
[Key]	nvarchar(512)	Kľúč stavu.
Value	nvarchar(512)	Hodnota stavu.

Tab. 25 – WerReportState

Názov stĺpca	Typ	Popis
Id	int	ID záznamu.
Report Id	int	Odkaz na WerReport.
[Key]	nvarchar(512)	Kľúč (napr. verzia operačného systému).
Value	nvarchar(512)	Hodnota.

Tab. 26 – WerReportOsInfo

Názov stĺpca	Typ údajov	Popis
Id	int	ID záznamu.
Report ID	int	Odkaz na WerReport.
Value	nvarchar(512)	Text zobrazený používateľovi.

Tab. 27 – WerReportUI

4.2.8 USN Journal (SchemaUsn)

Názov stĺpca	Typ údajov	Popis
Id	bigint	ID záznamu.
ArtObjectId	int	Prepojenie na ArtObject.
FileReferenceNumber	bigint	ID súboru (index MFT).
ParentFileReferenceNumber	bigint	ID nadradeného priečinka.
Usn	bigint	Číslo aktualizačnej sekvencie.
TimeStamp	datetime2(0)	Čas zmeny.
Reason	int	Dôvod zmeny (vlajky).
FileName	nvarchar(512)	Názov súboru.
FileAttributes	int	Atribúty súboru.

Tab. 28 – UsnRecordV2

Názov stĺpca	Typ	Popis
Id	bigint	ID záznamu.
FileReferenceNumber	binary(16)	128-bitové ID súboru.

ParentFileReferenceNumber	binárne (16)	128-bitové ID nadradeného súboru.
Usn	bigint	Sériové číslo.
TimeStamp	datetime2(0)	Čas zmeny.
Reason	int	Dôvod zmeny.
FileName	nvarchar(512)	Názov súboru.

Tab. 30 – UsnRecordV3

Názov stĺpca	Typ	Popis
Id	bigint	ID záznamu.
ParentFileReferenceNumber	binary(16)	128-bitové ID súboru.
Reason	int	Dôvod zmeny.
RemainingExtents	int	Zostávajúce rozsahy.
ExtentSize	smallint	Veľkosť rozsahu.

Tab. 31 – UsnRecordV4

4.2.9 Plánovač úloh systému Windows (SchemaWinScheduler)

Názov stĺpca	Typ údajov	Popis
Id	int	ID úlohy.
ArtObjectId	int	Prepojenie na súbor XML.
URI	varchar(512)	Cesta úlohy v systéme.
SecurityDescriptor	varchar(128)	Nastavenia zabezpečenia.
Author	nvarchar(128)	Autor úlohy.
Date	datetime2(0)	Dátum vytvorenia.
Description	nvarchar(1024)	Popis úlohy.
Enabled	bit	Či je úloha aktívna.
Hidden	bit	Či je úloha skrytá.

Tabuľka 32 – WinSchedulerTask

Názov stĺpca	Typ	Popis

Id	int	Hlavné ID.
TaskId	int	Príloha k úlohe.
UserId	varchar(256)	ID používateľa / SID.
LogonType	varchar(32)	Typ prihlásenia.
RunLevel	varchar(16)	Úroveň oprávnenia.

Tabuľka 33 – WinSchedulerTaskPrincipal

Názov stĺpca	Typ údajov	Popis
Id	int	ID zdieľania.
TaskId	int	Príloha k úlohe.
Command	nvarchar(512)	Príkaz / Cesta k EXE.
Arguments	nvarchar(MAX)	Parametre príkazu.
WorkingDirectory	nvarchar(512)	Pracovný adresár.

Tab. 34 – WinSchedulerTaskExecAction

Názov stĺpca	Typ údajov	Popis
Id	int	ID akcie.
TaskId	int	Prepojenie s úlohou.
Server	nvarchar(256)	SMTP server.
[To]	nvarchar(512)	Príjemca.
Subject	nvarchar(512)	Predmet.
Body	nvarchar(MAX)	Telo správy.

Tab. 35 – WinSchedulerSendEmailAction

Názov stĺpca	Typ údajov	Popis
Id	int	ID akcie.
TaskId	int	Prepojenie s úlohou.
ClassId	uniqueidentifier	CLSID objektu COM.
Data	xml	Dáta pre obslužnú funkciu.

Tabuľka 36 – WinSchedulerComHandlerAction

Názov stĺpca	Typ údajov	Popis
Id	int	ID akcie.
TaskId	int	Príloha k úlohe.
Title	nvarchar(256)	Názov okna.
Body	nvarchar(4000)	Text správy.

Tabuľka 37 – WinSchedulerShowMessageAction

Názov stĺpca	Typ údajov	Popis
Id	int	ID spúšťača.
TaskId	int	Prepojenie s úlohou.
Type	smallint	Typ spúšťača.
StartBoundary	datetime2(0)	Platné od.
EndBoundary	datetime2(0)	Expirácia.

Tab. 38 – WinSchedulerTaskTrigger

Názov stĺpca	Typ údajov	Popis
Id	int	ID záznamu.
TriggerId	int	Prepojenie s Trigger.
Delay	bigint	Oneskorenie (sekundy).

Tab. 39 – WinSchedulerBootTrigger

Názov stĺpca	Typ údajov	Popis
Id	int	ID záznamu.
TriggerId	int	Prepojenie s Trigger.
Subscription	nvarchar(256)	Predplatenie udalostí (dotaz).

Tabuľka 40 – WinSchedulerEventTrigger

Názov stĺpca	Typ údajov	Popis

Id	int	ID záznamu.
TriggerId	int	Prepojenie s Trigger.
UserId	nvarchar(256)	Používateľ, ktorého sa to týka.

Tab. 41 – WinSchedulerLogonTrigger

Názov stĺpca	Typ	Popis
Id	int	ID záznamu.
TriggerId	int	Pripojenie k spúšťaču.
RandomDelay	bigint	Náhodné oneskorenie.

Tab. 42 – WinSchedulerTimeTrigger

4.2.10 Činnosť časovej osi Windows 10 (SchemaWin10Timeline)

Názov stĺpca	Typ údajov	Popis
Id	int	ID aktivity.
AppId	varchar(1024)	JSON ID aplikácie.
ActivityType	int	Typ aktivity.
StartTime	datetime2(0)	Čas začatia.
EndTime	datetime2(0)	Čas ukončenia.
Payload	varbinary(1024)	Obsah aktivity.
PlatformDeviceId	varchar(128)	ID zariadenia.
ExpirationTime	datetime2(0)	Vypršanie platnosti záznamu.

Tab. 43 – Win10TimelineActivity

Názov stĺpca	Typ údajov	Popis
Id	int	ID operácie.
ArtObjectId	int	Odkaz na artefakt.
OperationOrder	int	Poradie operácie.
AppId	varchar(1024)	ID aplikácie.
Payload	varbinary(1024)	Operačné údaje.

Tab. 44 – Win10TimelineActivityOperation

Názov stĺpca	Typ údajov	Popis
Id	int	ID záznamu.
ActivityId	uniqueidentifier	ID aktivity.
PackageName	varchar(256)	Názov balíka.
Platform	varchar(256)	Platforma.

Tab. 45 – Win10TimelineActivityPackageId

Názov stĺpca	Typ údajov	Popis
Id	int	ID záznamu.
AssetPayload	varbinary(1024)	Binárny obsah.
LastRefreshTime	datetime2(0)	Čas poslednej aktualizácie.

Tab. 46 – Win10TimelineAsset (parser)

Názov stĺpca	Typ údajov	Popis
Id	int	ID záznamu.
AppId	varchar(256)	ID aplikácie.
AppTitle	varchar(256)	Názov aplikácie.

Tabuľka 47 – Win10TimelineAppSettings

Názov stĺpca	Typ údajov	Popis
Id	int	ID záznamu.
[Key]	varchar(128)	Kľúč metadát.
Value	varchar(128)	Hodnota.

Tab. 48 – Win10TimelineMetadata

4.2.11 Registre systému Windows (SchemaWinReg)

Názov stĺpca	Typ údajov	Popis
Id	bigint	ID kľúča (offset).

KeyName	nvarchar(255)	Názov kľúča.
LastWrittenTimestamp	datetime2(3)	Čas posledného zápisu.
Parent	int	ID nadradeného prvku.
NumberOfSubkeys	int	Počet podkľúčov.
NumberOfKeyValues	int	Počet hodnôt.

Tab. 49 – WinRegKeyNode

Názov stĺpca	Typ	Popis
Id	bigint	Hodnota ID.
CellId	bigint	ID kľúčovej bunky.
ValueName	nvarchar(MAX)	Názov hodnoty.
DataType	int	Typ údajov (REG_SZ atď.).
DataSize	int	Veľkosť dát.
DataOffset	int	Posun dát v súbore.

Tab. 50 – WinRegKeyValue

Názov stĺpca	Typ údajov	Popis
Id	bigint	ID bloku.
Signature	int	Podpis (regf).
LastWrittenTimestamp	datetime2(3)	Čas, kedy bol súbor naposledy upravený.
RootCellOffset	int	Posun koreňového kľúča.

Tab. 51 – WinRegBaseBlock

Názov stĺpca	Typ	Popis
Id	bigint	ID záznamu.
SecurityDescriptor	varbinary(2048)	Oprávnenia ACL (binárne).
ReferenceCount	int	Počet odkazov.

Tabuľka 52 – WinRegKeySecurity

Názov stĺpca	Typ údajov	Popis
Id	bigint	ID záznamu.
CellId	bigint	ID bunky.
ListSegmentOffset	int	Posun segmentu zoznamu.

Tab. 53 – WinRegBigData

Názov stĺpca	Typ údajov	Popis
Id	bigint	ID záznamu.
Offset	int	Pozícia v súbore.
Size	int	Veľkosť koša.

Tabuľka 54 – WinRegHiveBinHeader

Názov stĺpca	Typ	Popis
Id	bigint	ID záznamu.
Size	int	Veľkosť bunky.
Offset	int	Poloha bunky.

Tab. 55 – WinRegHiveBinCell

5 CTF dátové sady

Okrem súboru údajov Simulated Attackers' Techniques opísaného v predchádzajúcej časti sme ako zdroj údajov pre vytvorenie súborov údajov pre výskum digitálnej forenzie použili **súťaže Capture The Flag (CTF)**, hlavne preto, že prístup k reálnym údajom z bezpečnostných incidentov je často problematický a veľmi obmedzený. Dáta pochádzajúce zo skutočných kybernetických incidentov zvyčajne nie sú verejne dostupné a ich získanie je obmedzené organizačnými povoleniami, zákonnými a regulačnými požiadavkami a prísnymi povinnosťami v oblasti ochrany súkromia a údajov. Tieto obmedzenia výrazne bránia reprodukovateľnosti a experimentovaniu vo veľkom meradle v oblasti digitálnej forenznej analýzy.

Dataset CTF ponúka praktickú a eticky správnu alternatívu, pretože je špeciálne navrhnutý tak, aby emuloval realistické scenáre útokov a zároveň zostal legálne dostupný a dobre zdokumentovaný. Činnosti vykonávané počas výziev CTF sú zámerne škodlivé, ale vykonávajú sa v kontrolovanom prostredí, čo umožňuje bezpečný zber, zdieľanie a analýzu forenzných artefaktov. Výsledkom je, že dataset založený na CTF poskytuje cenné základné informácie, vďaka čomu je obzvlášť vhodný na školenie, testovanie a porovnávanie forenzných analytických metód a modelov strojového učenia.

Vytvorenie súboru údajov zo scenárov CTF prebiehalo podľa postupu opísaného vo výstupe D13 – Metóda vytvárania súboru údajov. Súbor údajov sa skladá z dvoch hlavných častí.

Prvá časť je **modifikovaná vkladaná dátová sada**, ktorá bola vytvorená pomocou nástroja *plaso* na generovanie superčasových osí, po čom nasledovala konštrukcia vkladanej reprezentácie extrahovaných forenzných údajov. Tento prístup umožňuje aplikovať pokročilé analytické a strojové učenie techniky na časové forenzné artefakty.

Druhou časťou je **modifikovaný súbor údajov EZ Tools**, ktorý bol generovaný pomocou forenzných nástrojov vyvinutých Ericom Zimmermanom na analýzu obrazov diskov a extrakciu forenzných artefaktov systému Windows. Extrahované údaje boli následne obohatené o štítky na podporu supervidovanej analýzy a hodnotenia.

Podrobné informácie o príprave dátových súborov z diskových obrazov CTF sú uvedené vo výstupe D13 – Metóda vytvárania dátových súborov.

5.1 Prípady použitia CTF

V rámci nášho súboru údajov sme vopred spracovali niekoľko reprezentatívnych forenzných prípadov pochádzajúcich zo simulovaných útokov a scenárov CTF (Capture The Flag) vykonaných v kontrolovanom prostredí. Všetky prípady sú založené na operačnom systéme Windows s súborovým systémom NTFS, ktorý zabezpečuje jednotnú štruktúru forenzných artefaktov a konzistentnú interpretáciu dôkazov. Predspracovanie údajov bolo navrhnuté tak, aby zjednotilo heterogénne forenzné zdroje, znížilo šum a zachovalo časové a kontextové

vzťahy medzi jednotlivými udalosťami. Nasledujúce prípady predstavujú rôzne scenáre útokov a situácií po incidente, ktoré sa používajú pri vytváraní a hodnotení súboru údajov.

V našom súbore údajov sme predspracovali nasledujúce prípady:

- Prípad CTF The Stolen Szechuan Sauce,
- Prípad Magnet CTF 2019,
- Prípad Magnet CTF 2022 a
- Prípad úniku údajov NIST.

5.1.1 Ukradnutá sečuánska omáčka

Prípad CTF „The Stolen Szechuan Sauce“ [8] sa zameriava na forenznú analýzu scenára porušenia bezpečnosti firemných údajov. Hlavným cieľom je zistiť, ako sa tajný recept na sečuánsku omáčku patriaci spoločnosti CITADEL dostal na temnú webovú stránku. Spoločnosť požiadala o forenznú analýzu svojho doménového radiča a sieťového hostiteľa s cieľom identifikovať škodlivé aplikácie nainštalované v systéme a určiť miesto a čas inštalácie softvéru. Prípad nám tiež poskytuje informácie o tom, či boli vytvorené, upravené alebo odstránené akékoľvek informácie a či došlo k porušeniu bezpečnosti údajov. Pracujeme s artefaktmi z doménového kontroléra spoločnosti (**DC server**) a z **pracovnej plochy** (sieťový hostiteľ).

5.1.2 Magnet CTF 2019 a 2022

Prípady Magnet CTF 2019 [9], a **Magnet CTF 2022** [10] sú scenáre forenzných výziev navrhnuté na testovanie špecifických zručností v oblasti digitálnej forenznej analýzy. Všetky tri prípady boli súčasťou súťaže Capture the Flag. Nejedná sa o klasický proces digitálnej forenznej analýzy, ale skôr o odpovede na konkrétne otázky, ako napríklad „kedy bol získaný obraz disku“, „kedy bol nainštalovaný softvér“ a podobne.

5.1.3 Prípad úniku dát NIST / Data Leakage Case

Prípad úniku dát NIST [11] predstavuje forenzný scenár zameraný na školenie, ktorý sa zameriava na vyšetrovanie rôznych foriem úniku dát. Hlavným cieľom tohto scenára je oboznámiť sa s rôznymi typmi exfiltrácie dát a precvičiť si vhodné forenzné vyšetrovacie techniky. Prípad opisuje vnútorné porušenie bezpečnosti dát, pri ktorom zamestnanec úmyselne zneužil svoj autorizovaný prístup na prenos dôverných informácií konkurenčnej spoločnosti. K úniku došlo kombináciou e-mailovej komunikácie, osobných cloudových úložísk a pokusu o fyzický prenos údajov pomocou externých pamäťových médií. Napriek existencii zavedených bezpečnostných politík a nasadených mechanizmov DLP/DRM sa podozrivý pokúsil obísť tieto kontroly, čo si vyžiadalo forenznú analýzu na identifikáciu stop, úniku údajov a rekonštrukciu činností vykonaných na zabavených elektronických zariadeniach.

5.2 Modifikovaný embedding súbor údajov

Aby sa zabezpečila **reprodukovateľnosť experimentov**, konzistentné spracovanie údajov a opätovná použiteľnosť vyškolených komponentov, všetky pomocné artefakty vytvorené počas prípravy súboru údajov boli uložené v štandardizovaných a široko podporovaných formátoch.

Štruktúra tejto časti súboru údajov je nasledovná:

- **global_scaler.pkl** – súbor pickle obsahujúci prispôsobený škálovač, ktorý sa používa počas inferencie na úpravu veľkosti delta hodnôt.
- **tokenizer** – adresár obsahujúci súbory potrebné na rekonštrukciu tokenizéra.
    - **merges** – textový súbor špecifikujúci, ako zlúčiť jednotlivé výstupy slovnej zásoby.
    - **vocab** – naučená slovná zásoba používaná na reprezentáciu textu v súboroch CTF.
- **processed_windows** – adresár obsahujúci predspracované okná (po 20 riadkov) vytvorené zo súborov CTF, formátované ako <embedding, delta, label>.
    - **val** – podadresár obsahujúci okná používané na validáciu počas trénovania.
    - **train** – podadresár obsahujúci okná použité na trénovanie.
    - ***_windows.pt** – okná uložené vo formáte optimalizovanom pre PyTorch (v dátovom súbore sú poskytované ako archívy: *_windows.pt.zip).
    - ***_windows.txt** – okná v textovom formáte čitateľnom pre človeka.

Slovník a tokenizér používané na spracovanie textového znázornenia forenzných artefaktov boli serializované a uložené vo **formáte JSON**. Tento formát poskytuje explicitné a transparentné mapovanie medzi tokenmi a ich numerickými identifikátormi, je čitateľný pre človeka a dá sa ľahko prenášať medzi rôznymi experimentálnymi prostredím, ľahko sa načíta do štandardných rámcov (napríklad tokenizéry od Hugging Face). Uloženie tokenizátora ako JSON zaručuje, že počas fázy trénovania aj hodnotenia sa uplatňujú identické pravidlá tokenizácie a slovná zásoba, čím sa zabraňuje nekonzistenciám v predspracovaní textu, ktoré by mohli negatívne ovplyvniť výkon modelu.

Na normalizáciu numerických vlastností, najmä časových rozdielov medzi po sebe idúcimi udalosťami, bol použitý mechanizmus škálovania (napr. MinMaxScaler). Po prispôsobení bol škálovač uložený ako **súbor pickle**, čo umožnilo jeho opätovné načítanie a konzistentné použitie na validáciu alebo novo získané údaje bez prepočítavania parametrov škálovania. Uloženie škálovacieho mechanizmu je nevyhnutné na udržanie jednotných numerických transformácií v celom experimentálnom procese a na zabezpečenie porovnateľnosti medzi modelmi trénovanými na rôznych častiach dátového súboru.

Namiesto ukladania samotných vložiek (ktoré by boli veľké) sa kódované okná ukladajú vo formáte optimalizovanom pre PyTorch.

V Tab. 56 sú uvedené štatistiky tejto časti dátového súboru.

Súbor		Veľkosť dátového súboru [kB]	typ
vocab	tokenizer	786	JSON
merges	tokenizer	446	TXT
Magnet_CTF_2019_windows	processed windows - train	7 278 650	PyTorch model file
Magnet_CTF_2019_windows	processed windows - train	1 071 261	TXT
NIST_Data_Leakage_windows	processed windows - train	6 883 445	PyTorch model file
NIST_Data_Leakage_windows	processed windows - train	936 726	TXT
SSS_DC_windows	processed windows - train	2 422 781	PyTorch model file
SSS_DC_windows	processed windows - train	362 009	TXT
SSS_Desktop_windows	processed windows - train	1 698 838	PyTorch model file
SSS_Desktop_windows	processed windows - train	236 472	TXT
Magnet_CTF_2022_windows	processed windows - val	8 350 791	PyTorch model file
Magnet_CTF_2022_windows	processed windows - val	1 173 086	TXT
global_scaler.pkl		0,609	Python pickle file

Tab. 56 – Veľkosť súborov

5.3 Modifikovaný súbor EZ Tools

Modifikovaná sada dát EZ Tools bola vytvorená pomocou súboru forenzných nástrojov vyvinutých Ericom Zimmermanom, uznávaným odborníkom v oblasti DFIR a autorom široko používaných nástrojov, ako sú RECmd, EvtxECmd, SrumECmd, PECmd, SBECmd a ďalšie. Tieto nástroje umožňujú štruktúrované parsovanie širokej škály forenzných artefaktov z operačného systému Windows, buď priamo z diskových obrazov, alebo z extrahovaných digitálnych stôp. Ďalšie podrobnosti o nástrojoch a ich použití na vytvorenie súboru údajov sú uvedené vo výstupe D13 – Metóda na vytvorenie súboru údajov.

Výsledný súbor údajov obsahuje normalizované a strojovo spracovateľné reprezentácie nízkolevelových systémových artefaktov, exportovaných predovšetkým vo formáte CSV, ktoré boli následne obohatené o štítky na podporu analytických úloh a úloh strojového učenia.

Dataset je usporiadaný do logických kategórií reprezentujúcich rôzne aspekty správania systému a používateľov. Každá kategória zodpovedá špecifickej skupine forenzných artefaktov

systému Windows a poskytuje doplnkové pohľady na činnosť systému, správanie používateľov a potenciálne akcie útočníkov.

Štruktúra tejto časti súboru údajov je nasledovná:

- Prípad použitia CTF (Magnet CTF 2019 / Magnet CTF 2022 / Prípad úniku údajov NIST / Ukradnutá omáčka Szechuan DC / Ukradnutá omáčka Szechuan Desktop) adresár:
    - Adresár EventLogs,
    - Adresár FileFolderAccess,
    - Adresár FileSystem,
    - ProgramExecution,
    - Adresár Registry,
    - adresár SRUMDatabase.

Podrobný prehľad každej kategórie je uvedený v nasledujúcich podkapitolách. Komplexný súhrn všetkých artefaktov a súvisiacich informácií je k dispozícii vo výstupe D12 – Metóda automatizovaného zberu digitálnych stôp.

5.3.1 EventLogs

Táto časť súboru údajov obsahuje analyzované systémové protokoly Windows (súbory .evtx) spracované pomocou EvtxECmd. EventLogs predstavujú jeden z najdôležitejších zdrojov časových a kontextových informácií v digitálnych forenzných vyšetrovaniach.

Dataset obsahuje systémové, bezpečnostné a aplikačné protokoly, ako aj rozšírené zdroje protokolovania, ako je Sysmon, ak sú k dispozícii. Tieto artefakty zaznamenávajú udalosti súvisiace s overovaním používateľov (prihlásenia a odhlásenia), využívaním oprávnení, činnosťou služieb a ovládačov, vytváraním a ukončovaním procesov, sieťovými pripojeniami, zmenami registrov a chybovými stavmi.

Kľúčové extrahované informácie zahŕňajú presné časové značky, identifikátory udalostí (Event ID), zdroje udalostí, popisné správy a štruktúrované dátové polia. Táto kategória je nevyhnutná na rekonštrukciu časových osí, detekciu podozrivého správania a koreláciu systémových udalostí s inými forenznými artefaktmi.

5.3.2 FileFolderAccess

Kategória FileFolderAccess sa zameriava na forenzné artefakty súvisiace s interakciami používateľov so súbormi a adresármi a poskytuje informácie o tom, ako používatelia navigovali v systéme a ku ktorým súborom alebo priečinkom mali prístup.

Primárne zdroje údajov zahŕňajú Shellbags (analyzované pomocou SBECmd alebo doplnkov registru), Jump Lists (JLECmd), položky RecentDocs, WordWheelQuery (história vyhľadávania v Průzkumníku Windows) a zoznamy MRU Open/Save. Spolu tieto artefakty odhaľujú

informácie o otvorených súboroch a priečinkoch, sieťových zdieľaných priečinkoch, používaní USB zariadení a prístupe k službám synchronizácie v cloude, ako sú OneDrive alebo Dropbox.

Táto kategória je kľúčová pre identifikáciu zámerov používateľov, rekonštrukciu pracovných postupov a detekciu pokusov o vyhľadávanie, prístup alebo exfiltráciu citlivých údajov.

5.3.3 Súborový systém

Kategória FileSystem zachytáva činnosť súborového systému na nízkej úrovni, ktorá pochádza predovšetkým z NTFS $MFT (Master File Table) analyzovaného pomocou MFTECmd. $MFT poskytuje komplexný prehľad metadát súborov a adresárov na celom diskovom objeme.

Súbor údajov obsahuje časové značky súborov, ako napríklad vytvorenie, úprava, prístup a zmena metadát (časy MACB), ako aj názvy súborov, cesty, veľkosti a príznaky označujúce odstránené alebo rezidentné súbory. Tieto informácie umožňujú podrobnú rekonštrukciu životného cyklu súboru, vrátane vytvorenia, úpravy, odstránenia a potenciálnej manipulácie.

Doplňujúce artefakty, ako sú skratky LNK (analyzované pomocou LECmd), umožňujú koreláciu aktivity súborového systému s interakciami používateľov a vykonávaním programov.

5.3.4 ProgramExecution

Táto kategória agreguje artefakty, ktoré poskytujú dôkazy o vykonávaní programov v systéme a sú kritické pre identifikáciu toho, ktoré aplikácie alebo binárne súbory boli spustené, ako často a v akom kontexte.

Zdroje údajov zahŕňajú súbory Prefetch (analyzované pomocou PECmd), Amcache, AppCompatCache (Shimcache), položky registra UserAssist, Background Activity Monitor (BAM/DAM) a Jump Lists. Spolu tieto artefakty odhaľujú spustené spustiteľné súbory, počty spustení, časy posledného spustenia, argumenty príkazového riadku a v niektorých prípadoch aj mechanizmy perzistencie.

Kategória ProgramExecution je obzvlášť relevantná pre detekciu spustenia škodlivého kódu, nástrojov na laterálny pohyb, binárnych súborov living-off-the-land (LOLBins) a neoprávneného používania softvéru.

5.3.5 Registre

Kategória Registry pozostáva z artefaktov analyzovaných z registrov Windows pomocou RECmd alebo Registry Explorer so špecializovanými doplnkami. Registre Windows slúžia ako centrálne úložisko konfigurácií systému, nastavení používateľov a stavov aplikácií, čo z nich robí bohatý zdroj forenzných stôp.

Súbor údajov obsahuje artefakty, ako sú UserAssist, TypedURLs, zoznamy MRU, kľúče automatického spustenia (Run a RunOnce), Shellbags, história USB zariadení (USBSTOR), konfigurácie služieb, nainštalované programy, mechanizmy perzistencie a údaje súvisiace s prehliadačom. Na extrakciu konkrétnych artefaktov, ako sú Bam, RecentApps alebo TypedPaths, bolo použitých viacero doplnkov.

Artefakty registrov sú nevyhnutné na pochopenie zmien konfigurácie systému, identifikáciu perzistencie a sledovanie histórie aktivít používateľov a používania periférnych zariadení.

5.3.6 SRUMDatabase

Kategória SRUMDatabase je odvodená z databázy System Resource Usage Monitor (SRUDB.dat), ktorá bola analyzovaná pomocou SrumECmd. Táto databáza ukladá historické štatistiky využitia systémových zdrojov a je mimoriadne cenná pre dlhodobú rekonštrukciu aktivít.

Táto časť súboru údajov obsahuje podrobné informácie o využívaní aplikácií (čas CPU, prečítané a zapísané bajty), sieťovej aktivite (odoslané a prijaté dáta podľa aplikácie a sieťového rozhrania), spotrebe energie a udalostiach, ako sú push notifikácie. Údaje SRUM sa zvyčajne uchovávajú približne 30–60 dní, čo umožňuje analýzu aktivít aj v prípade, že ostatné forenzné artefakty boli vymazané alebo prepísané.

Táto kategória poskytuje jedinečný kvantitatívny pohľad na správanie aplikácií a sieťovú aktivitu a je vysoko efektívna pri identifikácii anomálnych alebo podozrivých vzorov v dlhších časových obdobiach.

5.4 Charakteristiky modifikovaného súboru údajov EZ nástrojov

Táto časť sa zameriava na hodnoty a charakteristiky údajov extrahovaných z dátových súborov vytvorených pomocou nástrojov EZ. Poskytuje prehľad kľúčových štatistických vlastností artefaktov získaných z forenzných obrazov z prípadov použitia CTF.

5.4.1 Szechuan Sauce – DC

Nasledujúca tabuľka Tab. 57 zobrazuje štatistiky z artefaktov získaných z CTF Szechuan Sauce DC.

Artefakt		Veľkosť súboru údajov [kB]	Počet záznamov
EventLogs		44 200	87 280
FileDeletion		0,299	1
FileFolderAccess	AutomaticDestinations	9	9

	CustomDestinations	6	6
	LECmd_Output	7	12
	Administrator_NTUSER	0,256	1
	Administrator_UsrClass	5	27
FileSystem	MFTECmd_$MFT_Output	49 900	111 850
	MFTECmd_$J_Output	17 300	82 085
ProgramExecution	Amcache_UnassociatedFileEntries	1	2
	Windows81_Windows2012R2_SYSTEM_AppCompatCache	52	280
Registry	RECmd_Batch_AllRegExecutablesFoundOrRun_Output	19	29
	RECmd_Batch_BasicSystemInfo_Output	81	206
	RECmd_Batch_InstalledSoftware_Output	15	34
	RECmd_Batch_RECmd_Batch_MC_Output	6 300	17 372
	RECmd_Batch_RegistryASEPs_Output	16 700	42 987
	RECmd_Batch_SoftwareASEPs_Output	529	1 320
	RECmd_Batch_SoftwareClassesASEPs_Output	10 600	28 194
	RECmd_Batch_SoftwareWoW6432ASEPs_Output	6 600	17 038
	RECmd_Batch_SystemASEPs_Output	6 200	17 459
	RECmd_Batch_UserActivity_Output	107	198

Tab. 57 – Veľkosť súborov

5.4.2 Szechuan Sauce – Desktop

Nasledujúca tabuľka Tab. 58 zobrazuje štatistiky artefaktov získaných z CTF Szechuan Sauce Desktop.

Artefakt		Veľkosť súboru údajov [kB]	Počet záznamov
EventLogs		33 100	40 917
FileFolderAccess	AutomaticDestinations	51	47
	CustomDestinations	4	5
	LECmd_Output	18	28
	ricksanchez_UsrClass	1	6
	Administrator_UsrClass	7	34
	mortysmith_UsrClass	1	5
	Admin_UsrClass	0,822	3
FileSystem	MFTECmd_$MFT_Output	58 600	132 615
	MFTECmd_$J_Output	9 900	43 463

ProgramExecution	Amcache_DriverPackages	2	4
	Amcache_DeviceContainer	4	16
	Amcache_AssociatedFileEntries	37	83
	PECmd_Output	2 000	196
	Amcache_DriveBinaries	126	371
	Amcache_UnassociatedFileEntries	6	15
	Amcache_ShortCuts	6	35
	Amcache_ProgramEntries	56	85
	Windows10Creators_SYSTEM_AppCompatCache	52	266
	Amcache_DevicePnps	109	201
Registry	RECmd_Batch_AllRegExecutablesFoundOrRun_Output	109	194
	RECmd_Batch_BasicSystemInfo_Output	50	149
	RECmd_Batch_InstalledSoftware_Output	15	42
	RECmd_Batch_RECmd_Batch_MC_Output	6 400	16
	RECmd_Batch_SystemASEPs_Output	5 700	15 411
	RECmd_Batch_SoftwareASEPs_Output	955	2 758
	RECmd_Batch_RegistryASEPs_Output	26 600	77 615
	RECmd_Batch_UserClassesASEPs_Output	457	1 114
	RECmd_Batch_SoftwareClassesASEPs_Output	16 300	50 182
	RECmd_Batch_SoftwareWoW6432ASEPs_Output	6 900	20 921
	RECmd_Batch_UserActivity_Output	892	1 718
SRUMDatabase	SrumECmd_vfuprov_Output	19	105
	SrumECmd_NetworkUsages_Output	41	217
	SrumECmd_NetworkConnections_Output	2	10
	SrumECmd_AppTimelineProvider_Output	280	1488
	SrumECmd_PushNotifications_Output	2	11
	SrumECmd_AppResourceUseInfo_Output	169	766

Tab. 58 – Veľkosť súborov

5.4.3 Prípad úniku údajov NIST

Nasledujúca tabuľka Tab. 59 uvádza štatistiky z artefaktov získaných z prípadu úniku údajov CTF NIST.

Artefakt	Veľkosť súboru údajov [kB]	Počet záznamov
EventLogs	4 100	5 199

FileFolderAccess	AutomaticDestinations	39	42
	CustomDestinations	44	53
	LECmd_Output	23	36
	informant_UsrClass	19	97
	informant_NTUSER	3	12
	temporary_UsrClass	2	10
	admin11_UsrClass	1	6
FileSystem	MFTECmd_$MFT_Output	43 700	98 904
	MFTECmd_$J_Output	69 400	317 137
ProgramExecution	PECmd_Output	1 100	95
	Windows7x64_Windows2008R2_SYSTEM_AppCompatCache	51	262
	Windows7x64_Windows2008R2_SYSTEM_AppCompatCache	57	305
Registry	RECmd_Batch_AllRegExecutablesFoundOrRun_Output	80	127
	RECmd_Batch_BasicSystemInfo_Output	184	472
	RECmd_Batch_InstalledSoftware_Output	155	376
	RECmd_Batch_RECmd_Batch_MC_Output	12 300	33 161
	RECmd_Batch_SystemASEPs_Output	11 800	33 102
	RECmd_Batch_SoftwareASEPs_Output	718	1 865
	RECmd_Batch_RegistryASEPs_Output	35 400	92 418
	RECmd_Batch_UserClassesASEPs_Output	4	10
	RECmd_Batch_SoftwareClassesASEPs_Output	24 700	66 624
	RECmd_Batch_SoftwareWoW6432ASEPs_Output	14 400	37 640
	RECmd_Batch_UserActivity_Output	247	446

Tab. 59 – Veľkosť súborov

5.4.4 Magnet CTF 2019

Nasledujúca tabuľka Tab. 60 zobrazuje štatistiky artefaktov získaných z CTF Magnet 2019.

Artefakt		Veľkosť súboru údajov [kB]	Počet záznamov
EventLogs		93 300	109 433
FileDeletion		0,313	1
FileFolderAccess	AutomaticDestinations	32	30

	CustomDestinations	1	2
	LECmd_Output	14	24
	SelmaBouvier_UsrClass	3	16
	Administrator_UsrClass	6	28
FileSystem	MFTECmd_$MFT_Output	94 200	210 592
	MFTECmd_$J_Output	63 700	292 405
ProgramExecution	Amcache_DriverPackages	1	3
	Amcache_DeviceContainers	4	14
	Amcache_AssociatedFileEntries	9	21
	PECmd_Output	2 300	212
	Amcache_DriveBinaries	111	330
	Amcache_UnassociatedFileEntries	22	55
	Amcache_ShortCuts	11	65
	Amcache_ProgramEntries	59	104
	Windows10Creators_SYSTEM_AppCompatCache	152	742
	Amcache_DevicePnps	44	83
Registry	RECmd_Batch_AllRegExecutablesFoundOrRun_Output	71	130
	RECmd_Batch_BasicSystemInfo_Output	46	142
	RECmd_Batch_InstalledSoftware_Output	5	13
	RECmd_Batch_RECmd_Batch_MC_Output	5 600	14 732
	RECmd_Batch_SystemASEPs_Output	5 100	14 254
	RECmd_Batch_SoftwareASEPs_Output	835	2 434
	RECmd_Batch_RegistryASEPs_Output	24 500	72 291
	RECmd_Batch_UserClassesASEPs_Output	190	466
	RECmd_Batch_SoftwareClassesASEPs_Output	15 500	48 040
	RECmd_Batch_SoftwareWoW6432ASEPs_Output	6 900	20 900
	RECmd_Batch_UserActivity_Output	450	879
SRUMDatabase	SrumECmd_vfuprov_Output	108	569
	SrumECmd_NetworkUsages_Output	2 700	14 892
	SrumECmd_NetworkConnections_Output	204	1 374
	SrumECmd_AppTimelineProvider_Output	2 400	13 652
	SrumECmd_PushNotifications_Output	34	176
	SrumECmd_AppResourceUseInfo_Output	30 400	153
	SrumECmd_EnergyUsage_Output	2	12

Tab. 60 – Veľkosť súborov

5.4.5 Magnet CTF 2022

Nasledujúca tabuľka Tab. 61 uvádza štatistiky artefaktov získaných z CTF Magnet 2022.

Artefakt		Veľkosť súboru údajov [kB]	Počet záznamov
EventLogs		163 900	157 383
FileFolderAccess	AutomaticDestinations	50	56
	CustomDestinations	18	24
	LECmd_Output	32	55
	Patrick_usrClass	7	36
FileSystem	MFTECmd_$MFT_Output	198 200	418 398
	MFTECmd_$J_Output	74 400	344 746
ProgramExecution	Amcache_DriverPackages	27	30
	Amcache_DeviceContainers	2	7
	Amcache_AssociatedFileEntries	68	175
	PECmd_Output	8 400	786
	Amcache_DriveBinaries	142	391
	Amcache_UnassociatedFileEntries	74	188
	Amcache_ShortCuts		
	Amcache_ProgramEntries	75	120
	00_Windows10Creators_SYSTEM_AppCompatCache	101	498
	03_Windows10Creators_SYSTEM_AppCompatCache	120	554
	Amcache_DevicePnps	103	166
Registry	RECmd_Batch_AllRegExecutablesFoundOrRun_Output	122	234
	RECmd_Batch_BasicSystemInfo_Output	110	337
	RECmd_Batch_InstalledSoftware_Output	33	95
	RECmd_Batch_RECmd_Batch_MC_Output	10 500	31 093
	RECmd_Batch_SystemASEPs_Output	10 000	30 725
	RECmd_Batch_SoftwareASEPs_Output	2 000	5 702
	RECmd_Batch_RegistryASEPs_Output	57 200	165 565
	RECmd_Batch_UserClassesASEPs_Output	3 000	7 770
	RECmd_Batch_SoftwareClassesASEPs_Output	36 700	112 130
	RECmd_Batch_SoftwareWoW6432ASEPs_Output	13 800	41 203
	RECmd_Batch_UserActivity_Output	329	647

Tab. 61 – Veľkosť súborov

6 Bibliografia

[1] C. Grajeda, F. Breitinger, I. Baggili, Availability of datasets for digital forensics–and what is missing, Digital Investigation 22 (2017) S94–S105.

[2] L. Luciano, et al., Digital forensics in the next five years, in: Proceedings of the 13th International Conference on Availability, Reliability and Security, 2018.

[3] H. Studiawan, F. Sohel, C. Payne, Sentiment analysis in a forensic timeline with deep learning, IEEE Access 8 (2020) 60664–60675.

[4] E. Casey, The chequered past and risky future of digital forensics, Aust. J. Forensic Sci. 51 (6) (2019) 649–664.

[5] P. Sokol, et al., The analysis of digital evidence by Formal concept analysis, The 16th International Conference on Concept Lattices and Their Applications (CLA 2022), 2022.

[6] P. Sokol, et al., Formal concept analysis approach to understand digital evidence relationships, Int. J. Approx. Reason. 159 (2023) 108940.

[7] E. Marková, P. Sokol, K. Kováčová, Detection of relevant digital evidence in the forensic timelines, 2022 14th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), IEEE, 2022.

[8] Case 001 – the stolen szechuan sauce, 2020, [online] Dostupné na: https://dfirmadness.com/the-stolen-szechuan-sauce/.

[9] MagnetCTF 2019 Windows Desktop, 2019, [online] Dostupné na: https://digitalcorpora.s3.amazonaws.com/corpora/scenarios/magnet/2019%20CTF%20-%20Windows-Desktop.zip.

[10] Magnet CTF 2022 Windows, 2022, [online] Dostupné na: https://digitalcorpora.s3.amazonaws.com/corpora/scenarios/magnet/2022%20CTF%20-%20Windows.zip.

[11] Data LeakageCase, [online] Dostupné na: https://cfreds-archive.nist.gov/data_leakage_case/data-leakage-case.html.