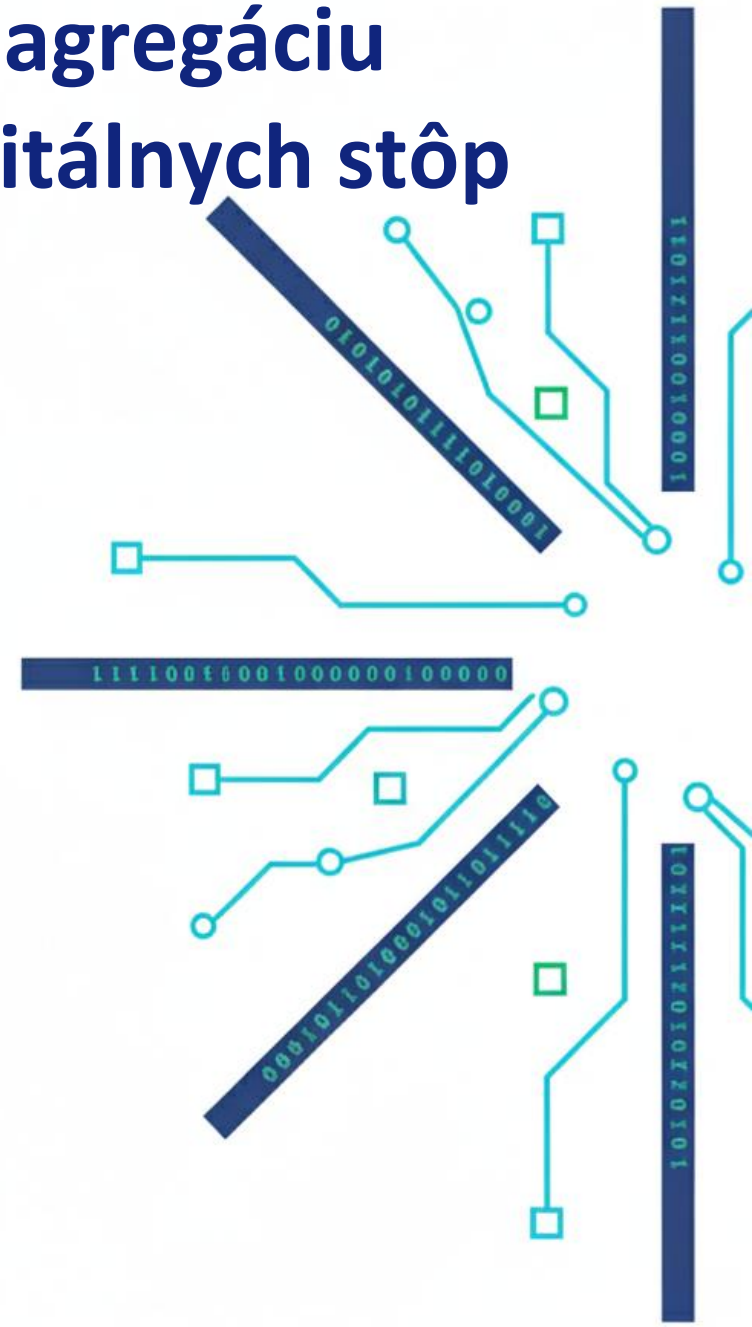
D17 – Model na agregáciu a prepojenie digitálnych stôp



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 3
2 Úvod 4
3 Agregácia forenzných artefaktov 4
3.1 Predspracovanie artefaktov 4
3.2 Identifikácia scenárov správania 5
3.3 Prepojenie digitálnych stôp 15
3.4 Agregácia digitálnych stôp 16
3.5 Vizualizácia agregovaných digitálnych stôp 21
4 Agregačné funkcie vzhľadom na neznáme hodnoty (NaN) 24
4.1 Hodnoty 0 a 1 sú rovnako významné 24
4.2 Hodnota 1 je viac významná ako 0 25
4.3 Hodnoty 0 a NaN sú rovnako významné 26
4.4 Zhrnutie a komparácia 27
5 Bibliografia 27

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 efektívne spracovanie veľkého množstva heterogénnych dát zásadnú výzvu, najmä pri rekonštrukcii udalostí a identifikácii kauzálnych súvislostí medzi jednotlivými artefaktmi. Moderné prístupy preto čoraz viac využívajú koncepty agregácie dát a spájania dát, ktoré umožňujú transformovať izolované forenzné stopy do analytikmi využiteľného modelu.

Tento dokument sa zameriava na návrh a implementáciu modelu agregácie a spájania forenzných stôp v rámci projektu ADFIR (Úloha KPB3.3 – Agregácia a spájanie digitálnych stôp). Cieľom je vytvoriť model, ktorý umožní systematické spracovanie artefaktov, redukciu množstva záznamov, ich koreláciu naprieč rôznymi zdrojmi a následnú interpretáciu v kontexte identifikovaných scenárov správania systému alebo používateľa.

Kľúčovou časťou práce je proces spájania forenzných stôp na základe spoločných identifikátorov, časových súvislostí a kontextových väzieb. Tento prístup umožňuje podporiť identifikované udalosti pomocou viacerých artefaktov a zároveň znížiť množstvo jednotlivých záznamov, ktoré musí analytik spracovať. Následne sú tieto spojené stopy agregované do definovaných časových okien, čím sa vytvára prehľadná časová os udalostí a umožňuje sa efektívnejšia analýza dynamiky sledovaných procesov a zmien.

Navrhovaný model tak poskytuje systematický spôsob, ako prejsť od fragmentovaných dát k integrovanému pohľadu na digitálne stopy, pričom podporuje nielen rekonštrukciu udalostí, ale aj identifikáciu vzorcov správania relevantných pre forenznú analýzu.

3 Agregácia forenzných artefaktov

Táto kapitola sa zameriava na systematické spracovanie digitálnych forenzných artefaktov s cieľom ich transformácie do analyticky využiteľnej podoby. Postupne pokrýva celý proces od predspracovania heterogénnych dát, cez identifikáciu scenárov správania až po prepojenie a agregáciu digitálnych stôp. Dôraz je kladený na vytvorenie jednotnej reprezentácie udalostí prostredníctvom metazáznamov, ktoré umožňujú efektívnejšiu interpretáciu a koreláciu údajov.

3.1 Predspracovanie artefaktov

V rámci spracovania forenzných dát boli využité viaceré zdroje artefaktov, ktoré sa líšia svojou štruktúrou, formátom aj úrovňou detailu. Tieto dáta boli reprezentované prevažne v tabuľkovej forme (napr. CSV výstupy z nástrojov), pričom jednotlivé artefakty obsahovali rôzne množiny atribútov a odlišné sémantické významy. Pred ich ďalším spracovaním bolo preto nevyhnutné realizovať fázu predspracovania, ktorej cieľom bolo zjednotenie dátovej reprezentácie, výber relevantných atribútov a odstránenie nekonzistencií. Spracovaniu týchto

dát sme sa podrobnejšie venovali vo výstupe D15 – Model na extrakciu digitálnych stôp do maticovej reprezentácie. V predmetnom výstupe sme sa zamerali na dva spôsoby predspracovania digitálnych forenzných artefaktov.

Ako prvý bol použitý vstupný dataset „Szechuan Sauce“ z portálu DFIR Madnes[1,2], ktorý bol spracovaný pomocou vybraných nástrojov od Erica Zimmermana [1]. Z obrazov diskov sme pomocou nasledovných nástrojov extrahovali digitálne stopy:

- EvtxECmd [2],
- MFTECmd [3],
- RECmd [4],
- PECmd [5],
- AppCompatCacheParser [6],
- JLECmd [7].

Výstupom spracovania boli štruktúrované CSV súbory reprezentujúce jednotlivé forenzné artefakty. V rámci ďalšieho spracovania boli z týchto artefaktov extrahované relevantné atribúty, ktoré slúžia ako základ pre následnú agregáciu dát.

Dataset „Szechuan Sauce“ spolu s ďalšími 4 datasetmi (výstup D14 – Syntetický dataset) boli použité aj v rámci návrhu agregačných funkcií nad dátami s neúplnými údajmi. Vstupným podkladom boli dáta získané prostredníctvom nástroja Plaso, pričom získané dáta najprv prechádzajú predspracovaním podľa výstupu D15 – Model na extrakciu digitálnych stôp do maticovej reprezentácie. Výstupom je jeden CSV súbor pre každý dataset.

3.2 Identifikácia scenárov správania

V rámci návrhu modelu agregácie a spájania forenzných stôp bola definovaná množina reprezentatívnych scenárov správania, ktoré zachytávajú typické a forenzne relevantné aktivity v analyzovanom systéme. Tieto scenáre správania predstavujú abstrahované udalosti vyššej úrovne, ktoré môžu byť popísané pomocou viacerých nízkoúrovňových artefaktov.

Pre účely tohto dokumentu označujeme tieto abstrahované udalosti ako metazáznamy. **Metazáznam** predstavuje logicky ucelenú jednotku informácie, ktorá vzniká spojením viacerých forenzných stôp na základe časovej, identitnej alebo kontextovej väzby. Cieľom metazáznamov je zjednodušiť interpretáciu dát a umožniť efektívnejšiu detekciu a rekonštrukciu incidentov.

[1] https://dfirmadness.com/case001/DC01-E01.zip
[2] https://dfirmadness.com/case001/DESKTOP-E01.zip

Proces definovania metazáznamov pozostával z identifikácie relevantných scenárov správania (napr. autentifikácia, vzdialený prístup, manipulácia so súbormi alebo perzistencia), ku ktorým boli následne priradené forenzné artefakty schopné indikovať výskyt danej udalosti. Pre každý metazáznam budú v ďalších častiach kapitoly špecifikované zdrojové artefakty, z ktorých je možné udalosť detegovať a kľúčové atribúty extrahované z týchto artefaktov. Niektoré z atribútov, ako napríklad parametre označujúce konkrétne zdrojové artefakty pre metazáznam, boli do jednotlivých artefaktov doplnené.

Na základe tejto metodiky boli definované nasledujúce metazáznamy:

- **Lokálne prihlásenie**

Atribút	Dátový typ	Zdrojový artefakt
EventRecordId	int	Event Logy
TimeCreated	datetime	Event Logy
EventId	int	Event Logy
Provider	string	Event Logy
ProcessId	int	Event Logy
ThreadId	int	Event Logy
Computer	string	Event Logy
ChunkNumber	int	Event Logy
MapDescription	string	Event Logy
UserName	string	Event Logy
RemoteHost	string	Event Logy
Target	string	Event Logy
Logon_Type	int	Event Logy
Logon_Id	int	Event Logy
Authentication_Package_Name	string	Event Logy
Logon_Process_Name	string	Event Logy
SourceFile	string	Event Logy
TargetUserSid	string	Event Logy
Created_Std Info	datetime	MFT
Created_File Name	datetime	MFT
File_Name	string	MFT
SI<FN	bool	MFT
Created_On	datetime	UserAccounts SAM
Last_Login_Time	datetime	UserAccounts SAM
User_Id	string	UserAccounts SAM
Groups	string	UserAccounts SAM
LastLoggedOnDisplayName	string	Registre
LastLoggedOnSAMUser	string	Registre
LastLoggedOnUser	string	Registre
LastLoggedOnUserSID	string	Registre

Last_Write_Timestamp	datetime	Registre
Timestamp	datetime	doplnené
src_EventLogs	bool	doplnené
src_MFT	bool	doplnené
src_UserAccounts_SAM	bool	doplnené
src_HKLM_Software_LogonUI	bool	doplnené
Ts_Event Logs	bool	doplnené
Ts_MFT_Std Info	bool	doplnené
Ts_MFT_File Name	bool	doplnené
Ts_SAM_Created	bool	doplnené
Ts_SAM_Last Login	bool	doplnené
Ts_Logon UI	bool	doplnené
Payload	string	Event Logy

- **Lokálne neúspešné prihlásenie**

Atribút	Dátový typ	Zdrojový artefakt
EventRecordId	int	Event Logs
TimeCreated	datetime	Event Logs
EventId	int	Event Logs
Provider	string	Event Logs
ProcessId	int	Event Logs
ThreadId	int	Event Logs
Computer	string	Event Logs
ChunkNumber	int	Event Logs
MapDescription	string	Event Logs
UserName	string	Event Logs
RemoteHost	string	Event Logs
Target	string	Event Logs
Logon Type	int	Event Logs
FailureReason2	string	Event Logs
SourceFile	string	Event Logs
Payload	json	Event Logs

- **RDP prihlásenie na cieľovom zariadení**

Atribút	Dátový typ	Zdrojový artefakt
EventRecordId	Integer	Event Logs
TimeCreated	Datetime	Event Logs
EventId	Integer	Event Logs
Provider	String	Event Logs
Channel	String	Event Logs

ProcessId	Integer	Event Logs
ThreadId	Integer	Event Logs
Computer	String	Event Logs
ChunkNumber	Integer	Event Logs
UserId	String	Event Logs
MapDescription	String	Event Logs
RemoteHost	String	Event Logs
Target	String	Event Logs
Logon Type	Integer	Event Logs
Logon Id	String	Event Logs
Authentication Package Name	String	Event Logs
Logon Process Name	String	Event Logs
Session ID	Integer	Event Logs
SourceFile	String	Event Logs
Payload	String	Event Logs

- **RDP neúspešné prihlásenie na cieľovom zariadení**

Atribút	Dátový typ	Zdrojový artefakt
EventRecordId	Integer	Event Logs
TimeCreated	Datetime	Event Logs
EventId	Integer	Event Logs
Provider	String	Event Logs
ProcessId	Integer	Event Logs
ThreadId	Integer	Event Logs
Computer	String	Event Logs
UserName	String	Event Logs
MapDescription	String	Event Logs
RemoteHost	String	Event Logs
SourceFile	String	Event Logs
Target User	String	Event Logs
Logon Type	Integer	Event Logs
Failure Reason	String	Event Logs
Payload	Json	Event Logs

- **RDP odhlásenie na cieľovom zariadení**

Atribút	Dátový typ	Zdrojový artefakt
RecordNumber	Integer	Event Logs
TimeCreated	Datetime	Event Logs
EventId	Integer	Event Logs
Provider	String	Event Logs
ProcessId	Integer	Event Logs

ThreadId	Integer	Event Logs
Computer	String	Event Logs
User Id	String	Event Logs
MapDescription	String	Event Logs
User Name	String	Event Logs
SourceFile	String	Event Logs
Target	String	Event Logs
Logon Type	Integer	Event Logs
Payload	Json	Event Logs

- **RDP prihlásenie na zdrojovom zariadení**

Atribút	Dátový typ	Zdrojový artefakt
EventRecordId	Integer	Event Logs
TimeCreated	Datetime	Event Logs
EventId	Integer	Event Logs
Provider	String	Event Logs
Channel	String	Event Logs
ProcessId	Integer	Event Logs
ThreadId	Integer	Event Logs
Computer	String	Event Logs
ChunkNumber	Integer	Event Logs
UserId	String	Event Logs
MapDescription	String	Event Logs
RemoteHost	String	Event Logs
SourceFile	String	Event Logs
Destination IP	String	Event Logs
Destination Name	String	Event Logs
Payload	Json	Event Logs

- **RDP odhlásenie na zdrojovom zariadení**

Atribút	Dátový typ	Zdrojový artefakt
TimeCreated	Datetime	Event Logs
EventId	Integer	Event Logs
Provider	String	Event Logs
ProcessId	Integer	Event Logs
ThreadId	Integer	Event Logs
Computer	String	Event Logs
User Id	String	Event Logs
MapDescription	String	Event Logs
SourceFile	String	Event Logs

Disconnect Reason	String	Event Logs
Payload	Json	Event Logs

- **Vytvorenie novej služby**

Atribút	Dátový typ	Zdrojový artefakt
Event Record Id	Integer	Event Logs
Time Created	Datetime	Event Logs
Event Id	Integer	Event Logs
Provider	String	Event Logs
Channel	String	Event Logs
Process Id	Integer	Event Logs
Thread Id	Integer	Event Logs
Computer	String	Event Logs
Chunk Number	Integer	Event Logs
User Id	String	Event Logs
Map Description	String	Event Logs
Executable Info	String	Event Logs
Source File	String	Event Logs
Service Name	String	Event Logs
Start Type	String	Event Logs
Account	String	Event Logs
Payload	String	Event Logs

- **Spustenie služby**

Atribút	Dátový typ	Zdrojový artefakt
Event Record Id	Int	Event Logs
Time Created	Datetime	Event Logs
Event Id	Int	Event Logs
Provider	string	Event Logs
Channel	string	Event Logs
Process Id	int	Event Logs
Thread Id	int	Event Logs
Computer	string	Event Logs
Map Description	string	Event Logs
Executable Info	string	Event Logs
Source File	string	Event Logs
Service Name	string	Event Logs
Status	string	Event Logs

- **Vytvorenie a/alebo spustenie naplánovanej úlohy**

Atribút	Dátový typ	Zdrojový artefakt
Description	String	Registre
Key Path	String	Registre
Hive Path	String	Registre
Last Write Timestamp	Datetime	Registre
Deleted	Bool	Registre
Time Created	Datetime	Registre, Event Logs
Last start	Datetime	Registre
Last stop	Datetime	Registre
Path	String	Registre
Command	String	Registre
EventRecordId	Integer	Event Logs
EventId	Integer	Event Logs
Provider	String	Event Logs
Channel	String	Event Logs
ProcessId	Integer	Event Logs
ThreadId	Integer	Event Logs
Computer	String	Event Logs
ChunkNumber	Integer	Event Logs
UserId	String	Event Logs
MapDescription	String	Event Logs
UserName	String	Event Logs
Task	String	Event Logs
Instance Id	String	Event Logs
SourceFile	String	Event Logs
Payload	Json	Event Logs

- **Spustenie procesu**

Atribút	Dátový typ	Zdrojový artefakt
Executable Name	string	Prefetch
Executable Path	string	Prefetch, ShimCache
Hash	string	Prefetch
Size	int	Prefetch
Run Count	int	Prefetch
Last Run	datetime	Prefetch
Previous Run 0	datetime	Prefetch
Previous Run 1	datetime	Prefetch
Previous Run 2	datetime	Prefetch
Previous Run 3	datetime	Prefetch
Previous Run 4	datetime	Prefetch

Previous Run 5	datetime	Prefetch
Previous Run 6	datetime	Prefetch
Volume0Name	string	Prefetch
Volume1Name	string	Prefetch
Directories	string	Prefetch
Files Loaded	string	Prefetch
Last Modified Time UTC	datetime	ShimCache

- **Prístup k súboru**

Atribút	Dátový typ	Zdrojový artefakt
Entry Number	int	MFT
Sequence Number MFT	int	MFT
Parent Path_MFT	string	MFT
Parent Entry Number	int	MFT
File Name	string	MFT, UsnJ
Extension_MFT	string	MFT
Has Ads	bool	MFT
Is Ads	bool	MFT
File Size	int	MFT
Created_MFT	datetime	MFT
SI_Created_MFT	bool	doplnené
FN_Created_MFT	bool	doplnené
Last Access_MFT	datetime	MFT
SI_Last Access_MFT	bool	doplnené
FN_Last Access_MFT	bool	doplnené
Last Modified_MFT	datetime	MFT
SI_Last Modified_MFT	bool	doplnené
FN_Last Modified_MFT	bool	doplnené
Last Record Change_MFT	datetime	MFT
SI_Last Record Change_MFT	bool	doplnené
FN_Last Record Change_MFT	bool	doplnené
Parent Path_J	string	UsnJ
Sequence Number J	int	UsnJ
Extension_J	string	UsnJ
Update Sequence Number	int	UsnJ
Update Reasons	string	UsnJ
File Attributes	string	UsnJ
Update Timestamp	datetime	UsnJ
App Id_Jump L	string	Jump List
Creation Time_Jump L	datetime	Jump List
Last Modified_Jump L	datetime	Jump List

Interaction Count_Jump L	int	Jump List
Target Created_Jump L	datetime	Jump List
Target Modified_Jump L	datetime	Jump List
Target Accessed_Jump L	datetime	Jump List
Path_Jump L	string	Jump List
File Size_Jump L	int	Jump List
Machine ID	string	Jump List

- **Vymazanie súboru**

Atribút	Dátový typ	Zdrojový artefakt
Delete Time	datetime	UsnJournal
File Name	string	UsnJournal
Extension	string	UsnJournal
Entry Number	int	UsnJournal
Sequence Number	int	UsnJournal
Parent Entry Number	int	UsnJournal
Update Reasons	string	UsnJournal

- **Stiahnutie súboru**

Atribút	Dátový typ	Zdrojový artefakt
Entry Number	int	MFT
Sequence Number	int	MFT
Parent Path	string	MFT
Parent Entry Number	int	MFT
In Use	bool	MFT
File Name	string	MFT
Extension	string	MFT
Has Ads	bool	MFT
File Size	int	MFT
Created0x10	datetime	MFT
Created0x30	datetime	MFT
LastAccess0x10	datetime	MFT
LastAccess0x30	datetime	MFT
Copied	bool	MFT
SI<FN	bool	MFT

- **Zmena časových atribútov súboru (timestomping)**

Atribút	Dátový typ	Zdrojový artefakt
Entry Number	int	MFT

Sequence Number	int	MFT
Parent Path	string	MFT
Parent Entry Number	int	MFT
In Use	bool	MFT
File Name	string	MFT
Extension	string	MFT
File Size	int	MFT
Created0x10	datetime	MFT
Created0x30	datetime	MFT
LastModified0x10	datetime	MFT
LastModified0x30	datetime	MFT
LastRecordChange0x10	datetime	MFT
LastRecordChange0x30	datetime	MFT
LastAccess0x10	datetime	MFT
LastAccess0x30	datetime	MFT

- Vznik spustiteľného súboru

Atribút	Dátový typ	Zdrojový artefakt
Entry Number	int	MFT, UsnJ
Parent Path	string	MFT, UsnJ
Parent Entry Number	int	MFT, UsnJ
Sequence Number	int	MFT
File Name	string	MFT, UsnJ
Extension	string	MFT, UsnJ
Has Ads	bool	MFT
File Size	int	MFT
Created0x10	datetime	MFT
Created0x30	datetime	MFT
Copied	bool	MFT
SI<FN	bool	MFT
Usn Create Time	datetime	UsnJ

- **Vznik archívu**

Atribút	Dátový typ	Zdrojový artefakt
Entry Number	int	MFT, UsnJ

Parent Path	string	MFT, UsnJ
Parent Entry Number	int	MFT, UsnJ
Sequence Number	int	MFT
File Name	string	MFT, UsnJ
Extension	string	MFT, UsnJ
Has Ads	bool	MFT
File Size	int	MFT
Created0x10	datetime	MFT
Created0x30	datetime	MFT
Copied	bool	MFT
SI<FN	bool	MFT
Usn Create Time	datetime	UsnJ

Takto definované metazáznamy tvoria základnú vrstvu nad agregovanými dátami a umožňujú prechod od izolovaných artefaktov k interpretovateľným udalostiam, ktoré sú priamo využiteľné pri forenznej analýze a detekcii kybernetických bezpečnostných incidentov.

3.3 Prepojenie digitálnych stôp

Po vytvorení metazáznamov vznikla situácia, v ktorej je jedna reálna udalosť reprezentovaná viacerými metazáznamami, pričom každý z nich je odvodený z odlišného artefaktu alebo zdroja dát. Tieto metazáznamy síce popisujú tú istú aktivitu, avšak líšia sa úrovňou detailu, presnosťou časových údajov a dostupnými atribútmi. Takáto redundancia je prirodzeným dôsledkom digitálnej forenznej analýzy založenej na heterogénnych dátach, no zároveň vytvára potrebu ich ďalšieho spracovania. Viacero atribútov metazáznamov, pre rôzne artefakty, ostáva nedefinovaná.

Z tohto dôvodu bol aplikovaný proces **spájania metazáznamov (data fusion)**, ktorého cieľom je zlúčiť viaceré čiastkové reprezentácie tej istej udalosti do jednej konzistentnej a komplexnej entity. Výsledkom tohto procesu je jednotná reprezentácia udalosti, ktorá integruje informácie z rôznych zdrojov a zároveň eliminuje duplicitné záznamy.

Spájanie metazáznamov je realizované na základe definovaných korelačných kritérií, ktoré zahŕňajú najmä:

- časovú blízkosť udalostí v rámci definovaného časového okna,
- zhodu identifikátorov (napr. používateľ, SID, názov zariadenia, IP adresa),
- kontextové väzby (napr. rovnaký proces, session ID, logon ID),
- typ udalosti (zhoda v kategórii metazáznamu).

Proces spájania metazáznamov pozostáva z identifikácie kandidátnych metazáznamov, ktoré pravdepodobne reprezentujú tú istú udalosť, a ich následného zlúčenia do jedného výsledného metazáznamu. Pri zlúčení dochádza k:

- výberu najpresnejších alebo najspoľahlivejších hodnôt atribútov,
- doplneniu chýbajúcich údajov z iných metazáznamov,
- zachovaniu informácie o pôvode dát.

Výsledok predstavuje obohatený metazáznam, ktorý poskytuje komplexnejší pohľad na analyzovanú udalosť a zároveň znižuje množstvo redundancie v dátach. Tento prístup umožňuje efektívnejšiu analýzu, presnejšiu rekonštrukciu priebehu udalostí a lepšiu interpretáciu správania systému alebo používateľa. Prepojenie digitálnych stôp tak predstavuje kľúčový krok medzi extrakciou jednotlivých forenzných stôp a ich finálnou agregáciou do časových alebo scenárových celkov, pričom zásadne zvyšuje kvalitu a využiteľnosť výsledného dátového modelu.

3.4 Agregácia digitálnych stôp

Agregácia metazáznamov predstavuje nadstavbovú fázu spracovania, v ktorej sú už vytvorené a prípadne zlúčené metazáznamy usporadúvané do väčších analytických celkov. Cieľom tejto fázy je znížiť fragmentáciu evidovaných udalostí, zvýrazniť ich vzájomné súvislosti a vytvoriť prehľadnejší obraz o priebehu analyzovaných aktivít. V digitálnej forenznej analýze je takýto prístup prirodzeným pokračovaním spracovania artefaktov, pretože umožňuje prejsť od jednotlivých udalostí k uceleným sekvenciám alebo scenárom správania.

Prvým spôsobom agregácie je zoskupovanie metazáznamov podľa zvoleného časového okna. Pri tomto prístupe sú udalosti, ktoré nastali v rovnakom alebo časovo blízkom intervale, spájané do jedného agregovaného celku. Takto vytvorené skupiny umožňujú sledovať lokálne zhluky aktivít, analyzovať ich časovú následnosť a vytvárať prehľadné časové úseky, ktoré môžu reprezentovať napríklad konkrétnu reláciu používateľa, fázu útoku alebo samostatný bezpečnostný incident.

Veľkosť časového okna je pritom možné prispôsobiť charakteru analyzovaných dát a požadovanej granularite výsledkov. Menšie časové okná (napr. 30 sekúnd) umožňujú jemnejšie zachytiť bezprostredne nadväzujúce udalosti, zatiaľ čo väčšie okná (napríklad 24 hodín) podporujú identifikáciu širších aktivít pozostávajúcich z viacerých navzájom súvisiacich krokov. Súčasťou agregácie môžu byť aj vybrané parametre, napríklad typ udalosti, identita používateľa, názov hostiteľa, proces, cesta k súboru alebo relácia, čím sa znižuje riziko, že budú do jedného celku nesprávne zlúčené časovo blízke, ale vecne nesúvisiace udalosti. Pri takejto agregácii, máme možnosť vybrať niekoľko parametrov, ale aj všetky, teda okrem času.

Druhým spôsobom agregácie je zoskupovanie metazáznamov podľa vybraného spoločného parametra, teda bez primárneho dôrazu na pevne definované časové okno. V tomto prípade

sa vytvárajú scenárové celky reprezentujúce aktivity viazané na konkrétnu entitu, ako je používateľský účet, súbor, proces, služba, IP adresa alebo zariadenie. Takýto spôsob agregácie umožňuje sledovať správanie konkrétneho objektu naprieč viacerými udalosťami a lepšie identifikovať jeho význam v analyzovanom deji.

Príkladom môže byť agregácia všetkých metazáznamov viazaných na konkrétneho používateľa, čím vznikne ucelený prehľad jeho prihlásení, spustených procesov, prístupov k súborom alebo vzdialených relácií. Podobne je možné vytvoriť scenárový celok pre konkrétny súbor, v rámci ktorého budú prepojené udalosti jeho vzniku, modifikácie, prístupu, stiahnutia, archivácie alebo vymazania. Výsledkom je analytická štruktúra, ktorá nepodáva len časový pohľad na udalosti, ale aj entitne orientovaný pohľad na ich priebeh a vzájomné väzby.

Oba prístupy agregácie sa navzájom dopĺňajú a umožňujú analyzovať dáta z dvoch rozdielnych, no komplementárnych perspektív. Časová agregácia podporuje rekonštrukciu deja v chronologickej forme, zatiaľ čo agregácia podľa entity pomáha budovať scenárové alebo objektové pohľady na správanie systému a používateľov. Tieto dva prístupy je samozrejme možné kombinovať, kedy sa nezameriavame len na vybrané entity, ale aj na konkrétne časové úseky, v ktorých udalosti nastali. V kombinácii tieto prístupy zvyšujú interpretačnú hodnotu výsledného modelu a poskytujú vhodný základ pre následnú forenznú analýzu, detekciu incidentov a rekonštrukciu komplexných činností v systéme.

Pre agregáciu metazáznamov sme zvolili časový prístup, pričom jednotlivé metazáznamy sme agregovali v rámci každého scenára správania zvlášť aj s využitím parametrov. Využívame vysokoúrovňovú agregáciu (v nasledujúcich tabuľkách „Agregácia 1“), ktorá zlučuje všetky udalosti rovnakého typu (napr. všetky lokálne prihlásenia) v požadovanom časovom okne, aj agregáciu druhej úrovne (v nasledujúcich tabuľkách „Agregácia 2“), ktorá už zohľadňuje vybrané parametre (napr. SID používateľa).

Pre jednotlivé scenáre sme zvolili nasledovné atribúty pre agregáciu:

- **Lokálne prihlásenie**

Časový atribút	**Udalosť**	**Agregácia 1**		**Agregácia 2**	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
Timestamp	Local log on from "Computer" to user "Target"	Target	X local log-ons to user "Target"	X	
Timestamp	First log on of user "UserName"	x		X	
Timestamp	Last logged on user was "UserName"	x		x	

Created_On	Creation of user "UserName"	Všetko	X users created.		

- **RDP prihlásenie na cieľovom zariadení**

Časový atribút	Udalosť	Agregácia 1		Agregácia 2	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
TimeCreated	RDP login from "RemoteHost" to "Target".	Všetko	X RDP log ins.	RemoteHost	X RDP log ins from "Remote Host".
TimeCreated		Všetko	X RDP log ins.	RemoteHost	X RDP log ins from "Remote Host".

- **RDP prihlásenie na zdrojovom zariadení**

Časový atribút	Udalosť	Agregácia 1		Agregácia 2	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
TimeCreated	RDP connection from "Computer" to "Destination Name".	Všetko	X RDP connections.	Destination Name	X RDP connections to "Destination Name"
TimeCreated	RDP connection from "Computer" to "Destination IP".	Všetko	X RDP connections.	Destination Name	X RDP connections to "Destination Name"

- **RDP odhlásenie na zdrojovom zariadení**

Časový atribút	Udalosť	Agregácia 1		Agregácia 2	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
TimeCreated	RDP logoff initiated by computer "Computer".	Všetko	X RDP logoffs initiated.	x	

- **Vytvorenie novej služby**

Časový atribút	Udalosť	Agregácia 1		Agregácia 2	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť

TimeCreated	Service "Service Name" with "Start type" start was installed by "User Id".	Všetko	X services were installed	User Id	X services were installed by "User Id" user

- **Spustenie služby**

Časový atribút	Udalosť	Agregácia 1		Agregácia 2	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
TimeCreated	Service "Service Name" has been launched.	Všetko	X services launched	x	

- **Vytvorenie a/alebo spustenie naplánovanej úlohy**

Časový atribút	Udalosť	Agregácia 1		Agregácia 2	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
TimeCreated	Scheduled task "Task" was created by "User Id" to run under "User Name" user.	Všetko	X scheduled tasks created.	User Id	X scheduled tasks created by user "User Id"
TimeCreated	Scheduled task "Task" associated with user "User Id" was created.	Všetko	X scheduled tasks created.	User Id	X scheduled tasks associated with "User Id" created
TimeCreated	Scheduled task "Task" started under user "User Name".	Všetko	X scheduled tasks starteded.	User Name	X scheduled tasks started under user "User Name"
TimeCreated	Scheduled task "Task" was triggered by logon of user "User Name".	Všetko	X scheduled tasks triggered by logon	User Name	X scheduled tasks triggerd by logon of "User Name"

- **Spustenie procesu**

Udalosť	Agregácia 1	Agregácia 2

Časový atribút		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
Last Run	"ExecutableName" process was executed from "ExecutablePath"	Všetko	X processes were executed	ExecutablePath	X processes were executed from "ExecutablePath" path.

- **Prístup k súboru**

Časový atribút	Udalosť	Agregácia 1		Agregácia 2	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
Last Access_MFT	Access to "FileName" file.	Všetko	X files accessed	x	
Target Accessed_Jump L	Access to "Path_Jump L" file.	Všetko	X files accessed	x	

- **Vymazanie súboru**

Časový atribút	Udalosť	Agregácia 1		Agregácia 2	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
Delete Time	Deletion of "FileName" file.	Všetko	X files were deleted	x	

- **Stiahnutie súboru**

Časový atribút	Udalosť	Agregácia 1		Agregácia 2	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
Created0x10	File "FileName" was downloaded.	Všetko	X files were downloaded	Extension	X files with "Extension" extension were downloaded.

- **Zmena časových atribútov súboru (timestomping)**

Časový atribút	Udalosť	Agregácia 1		Agregácia 2	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
LastModified0x10	Suspected timestomping of "FileName".	Všetko	Suspected timestomping of X files.	Extension	Suspected timestomping of X "Extension" files.

- **Vznik spustiteľného súboru**

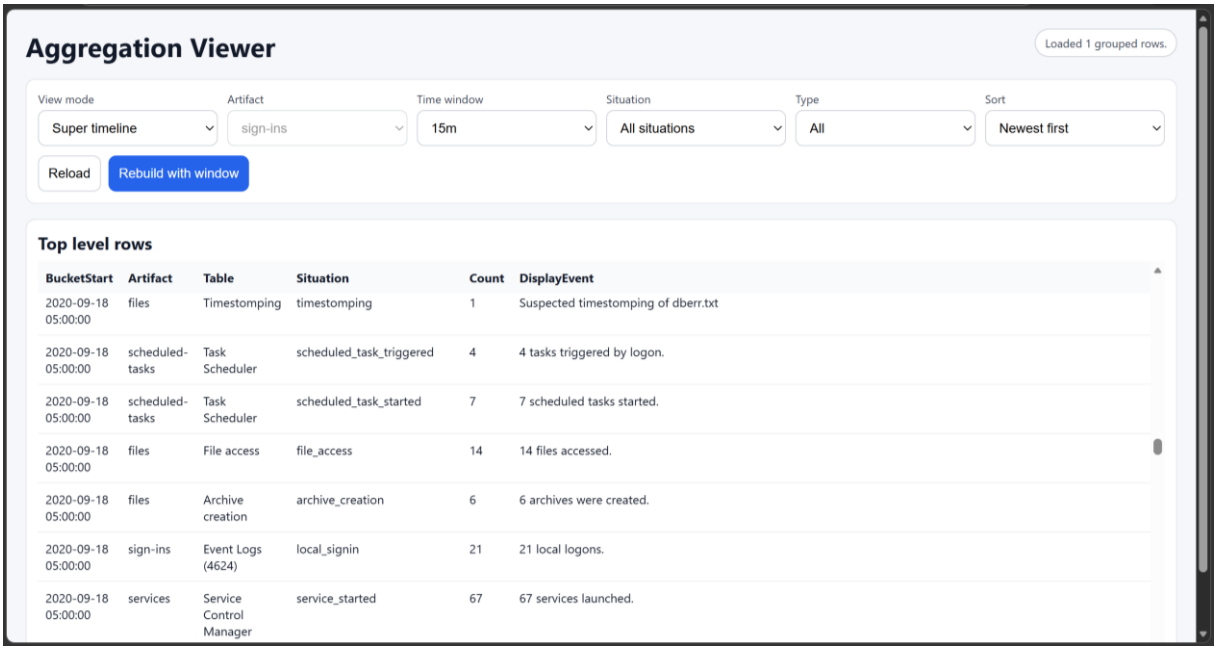
Časový atribút	Udalosť	Agregácia 1		Agregácia 2	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
Created0x10	Executable file "FileName" was created in "Parent Path".	Všetko	X executables were created	Parent Path	X execuatables were created in "ParentPath" folder

- **Vznik archívu**

Časový atribút	Udalosť	Agregácia 1		Agregácia 2	
		Agregačný atribút	Udalosť	Agregačný atribút	Udalosť
Created0x10	Archive "FileName" was created in path "ParentPath".	Všetko	X archives were created	Parent Path	X archives were created in "ParentPath" folder

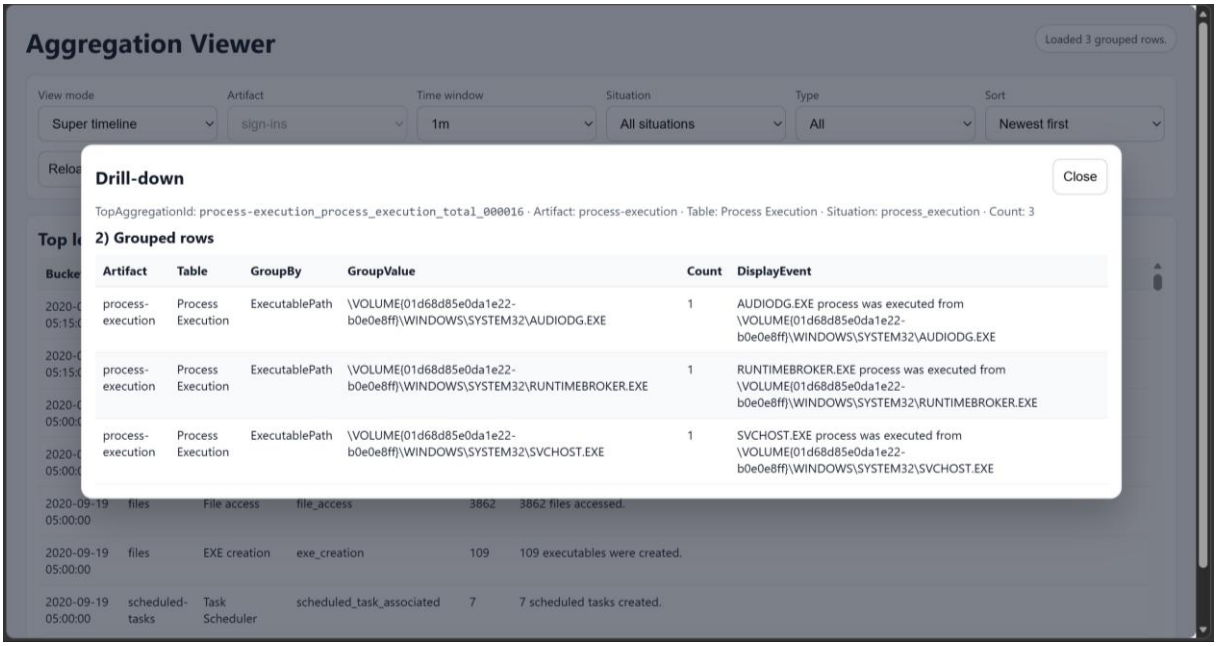
3.5 Vizualizácia agregovaných digitálnych stôp

Nad spracovanými metazáznamami bol vytvorený vizualizačný nástroj (Obrázok 1), ktorého úlohou je prehľadne zobrazovať agregované výsledky a uľahčiť ich ďalšiu analýzu. Do zobrazenia boli vybrané len najdôležitejšie parametre jednotlivých metazáznamov, aby sa zachoval dobrý prehľad a zároveň sa zachytila podstata jednotlivých udalostí. Nástroj podporuje nastavenie veľkosti agregačného okna, triedenie záznamov podľa času a filtračné mechanizmy, čo umožňuje prispôsobiť pohľad na dáta potrebné pre aktuálnu analytickú úlohu.

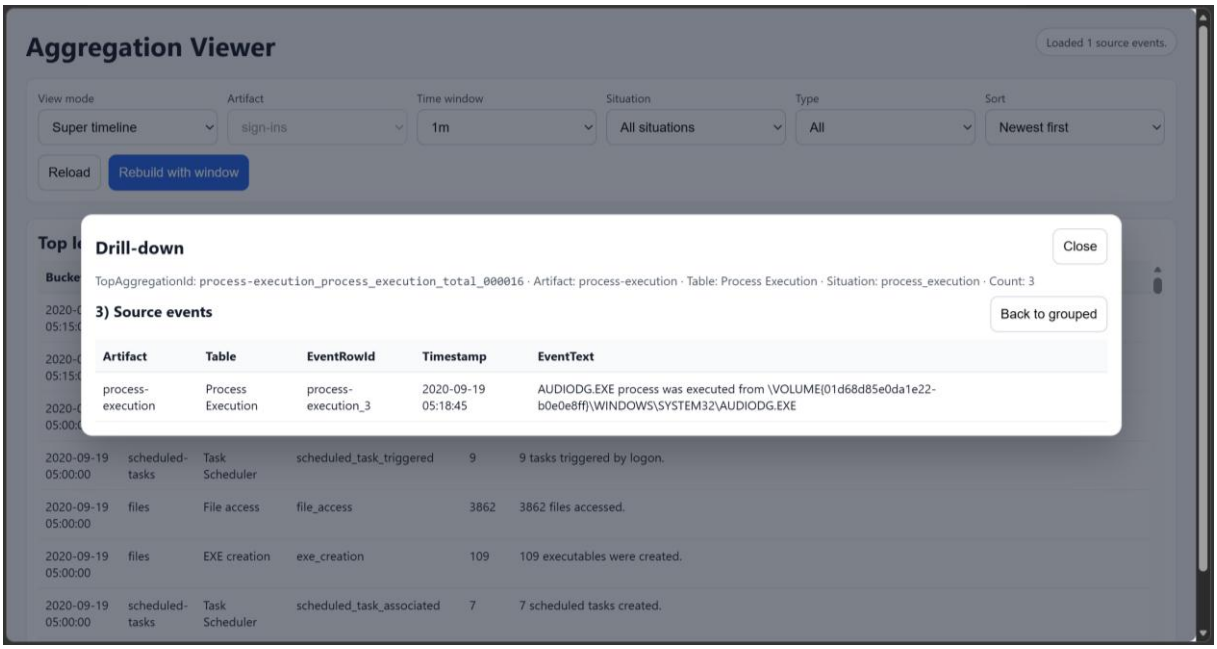


Obrázok 1 – Nástroj na vizualizáciu agregácii

Zobrazenie pracuje na úrovni agregovaných metazáznamov, pričom po zobrazení konkrétneho agregovaného záznamu je možné vidieť aj jednotlivé metazáznamy, ktoré boli do výslednej agregácie zahrnuté. Súčasťou riešenia je aj väzba na pôvodné záznamy v artefaktoch, vďaka čomu je možné spätne dohľadať zdroj informácie a overiť si, z akých dát bola výsledná udalosť odvodená (Obrázok 2 a Obrázok 3). Takýto prístup prepája vysokú mieru prehľadnosti s požiadavkou na dohľadateľnosť a transparentnosť forenzného spracovania dát.



Obrázok 2 – Väzba agregovaných údajov na pôvodné záznamy v artefaktoch – príklad 1



Obrázok 3 - Väzba agregovaných údajov na pôvodné záznamy v artefaktoch – príklad 2

4 Agregačné funkcie vzhľadom na neznáme hodnoty (NaN)

V rámci skúmania metodík spracovania dát s neúplnými údajmi (t. j. s výskytom chýbajúcich hodnôt NaN) sme sa zamerali aj na návrh a formálnu špecifikáciu **agregačných funkcií**, ktoré dokážu explicitne reflektovať **prítomnosť týchto neznámych údajov**. Z matematického hľadiska chápeme tieto funkcie ako zobrazenia $A\colon \mathrm{N}_0^3 \to [0,1]$, ktoré transformujú diskrétne početnosti (abundancie) výskytu stavov na spojitú mieru relevancie alebo hustoty informácie. Tradičné agregačné operátory ako minimum, maximum, súčet či aritmetický priemer sú koncipované tak, aby pracovali výhradne s dostupnými numerickými hodnotami. Informáciu o počte alebo podiele chýbajúcich hodnôt pritom nezohľadňujú, čo môže viesť k neúplnej alebo skreslenej interpretácii výsledkov. V mnohých dátových štruktúrach však práve množstvo chýbajúcich hodnôt predstavuje dôležitú informáciu (napr. v senzorových sieťach môže NaN signalizovať poruchu uzla, čo je kvalitatívne iný stav než nameraná nula). Z tohto dôvodu je vhodné definovať nové, resp. rozšíriť klasické agregácie spôsobom, ktorý umožní modifikovať ich výsledok v závislosti od neúplnosti údajov. Takto navrhnuté agregačné funkcie môžu výrazne prispieť k presnejšej interpretácii dát.

Nami navrhované agregačné funkcie pracujú s trojicou vstupných parametrov, ktoré sumarizujú základné charakteristiky binárnych dát: počet núl ($\#0$), počet jednotiek ($\#1$) a počet chýbajúcich hodnôt ($\#\mathrm{NaN}$). Dôležitým predpokladom je vždy väzba

$$\#0 + \#1 + \#\mathrm{NaN} = N,$$

kde N je prirodzené číslo. Táto väzba implikuje, že takýto údaj môžeme reprezentovať ako bod v diskrétnom simplexe v trojrozmernom priestore. Teda zníženie počtu 0 sa musí prejaviť v zvýšení počtu 1, resp. NaN hodnôt tak, aby ich súčet zostal zachovaný. Rovnako pre ostatné parametre. V závislosti od povahy spracovávaných dát a od toho, aký význam je pripisovaný jednotlivým hodnotám, rozlišujeme tri typové situácie uvedené nižšie. Tieto situácie zodpovedajú rôznym sémantickým interpretáciám neistoty v dátach. Každá z nich reprezentuje odlišný interpretačný rámec, a preto si vyžaduje špecificky definovanú agregačnú funkciu. Výsledná hodnota agregácie je určená požadovanými vlastnosťami, ktoré má funkcia spĺňať – najmä spôsobom, akým má reagovať na prítomnosť chýbajúcich údajov, na dominanciu jednej z binárnych hodnôt alebo na potrebu zachovať určitú monotónnosť či robustnosť voči neúplným dátam. V každom z uvažovaných prípadov uvažujeme agregačné funkcie, ktorých výstupom je hodnota z intervalu $[0,1]$, pričom hraničné hodnoty sa nadobúdajú. Tento interval predstavuje prirodzený rozsah pre interpretáciu agregácie, keďže zabezpečuje kompatibilitu s pravdepodobnostnou interpretáciou, normalizáciu naprieč rôznymi datasetmi (porovnávanie agregovaných hodnôt) a stabilitu výpočtov, čo je dôležité z technického hľadiska.

4.1 Hodnoty 0 a 1 sú rovnako významné

V situácii, keď sú hodnoty 0 a 1 považované za rovnako významné, je nutné, aby agregácia bola symetrická vzhľadom na tieto dve kategórie. Symetria (zámennosť) znamená, že

agregácia nerozlišuje medzi týmito hodnotami, obe patria do rovnakej množiny platných binárnych hodnôt. Formálne požadujeme, aby platilo $A(\#0, \#1, \#\mathrm{NaN}) = A(\#1, \#0, \#\mathrm{NaN})$. Agregačná funkcia teda má reagovať iba na ich spoločný počet vzhľadom na počet neznámych hodnôt NaN. V tomto nastavení sa prirodzene ukazuje, že:

- maximum agregácie sa dosiahne vtedy, keď sú všetky údaje známe (t. j. pozostávajú výlučne z 0 a 1),
- minimum nastáva v prípade, keď údaje pozostávajú z neznámych hodnôt (samé NaN), pretože agregácia nemá k dispozícii žiadnu informáciu.

Na základe týchto požiadaviek je vhodnou voľbou nasledujúca agregačná funkcia:

$$A(\#0, \#1, \#\mathrm{NaN}) = \frac{\#0 + \#1}{\#0 + \#1 + \#\mathrm{NaN}}$$

Táto funkcia vyjadruje podiel známych binárnych hodnôt vzhľadom na celkový počet hodnôt (vrátane neznámych). Predstavuje mieru „kompletnosti“ dátového súboru alebo možno povedať, že vyjadruje šancu, že náhodne vybraný prvok zo súboru nie je chýbajúci. Agregačná funkcia taktiež svojou konštrukciou spĺňa vlastnosť monotónnosti (pri zachovaní konštantného N):

- pri fixnom počte neznámych NaN je agregácia konštantná vzhľadom na počet 0 a 1, keďže ich nerozlišuje (záleží iba na ich súčte),
- naopak, agregačná funkcia je klesajúca vzhľadom na rastúci počet NaN: čím viac neznámych údajov, tým menší podiel platných informácií.

4.2 Hodnota 1 je viac významná ako 0

V tomto prípade je dôležitý predpoklad, že výskyt hodnoty 1 má vyššiu informačnú hodnotu než výskyt hodnoty 0. Z tohto dôvodu priraďujeme jednotke váhu 1, zatiaľ čo nule priraďujeme váhu $w \in (0,1)$, ktorá vyjadruje jej relatívne nižší význam. Je prirodzené požadovať, aby zvyšujúci sa počet neznámych hodnôt NaN viedol k systematickému poklesu agregovanej hodnoty, keďže dostupná informácia predstavuje čoraz menší podiel z celkového počtu známych hodnôt. Na základe to by agregačná funkcia mala spĺňať, že

- maximálna hodnota agregácie sa nadobúda v prípade, že počet hodnôt 1 je maximálny, a teda počet 0 aj NaN nulový,
- minimálna hodnota agregácie sa dosiahne v prípade, keď všetky hodnoty sú neznáme, teda NaN,
- ak sú všetky hodnoty 0, agregácia nadobúda hodnotu práve w, čo odráža zvolenú váhu pre hodnotu 0.

Na základe uvedeného navrhujeme nasledujúcu agregačnú funkciu:

$$A(\#0, \#1, \#\mathrm{NaN}) = \frac{\#1 + w \cdot \#0}{\#0 + \#1 + \#\mathrm{NaN}}, \quad w \in (0,1)$$

Takto definovaná agregačná funkcia rešpektuje asymetriu významu medzi hodnotami 1 a 0, pričom zároveň citlivo reaguje na prítomnosť chýbajúcich údajov a zachováva interpretovateľnosť výsledku v normalizovanom intervale $[0,1]$. Presnejšie, parametrizácia pomocou w dovoľuje jemne ladiť citlivosť modelu na prítomnosť „negatívnych“ (0) oproti „pozitívnym“ (1) výsledkom. Zároveň sú splnené vlastnosti monotónnosti (pri fixnom N), teda navrhovaná agregačná funkcia

- je rastúca vzhľadom na rastúci počet 1 pri fixnom počte 0 alebo NaN, keďže jednotky majú najvyššiu váhu,
- je rastúca vzhľadom na rastúci počet 0 pri fixnom počte 1, pretože nuly prispievajú do čitateľa kladnou váhou,
- je klesajúca vzhľadom na rastúci počet 0 pri fixnom počte NaN, keďže nárast hodnôt 0 zároveň zmenšuje počet hodnôt 1,
- je klesajúca vzhľadom na rastúci počet NaN, keďže nárast neznámych hodnôt zmenšuje počet hodnôt 0 a 1.

4.3 Hodnoty 0 a NaN sú rovnako významné

V prípade, že hodnota 1 je jediným nositeľom relevantnej informácie a hodnoty 0, NaN majú z hľadiska interpretácie rovnaký význam (symetrickosť), je potrebné agregačnou funkciou zachytiť výlučne relatívny podiel hodnôt 1 v celom súbore hodnôt. Táto situácia je typická pre riedke dáta (sparse data), kde nás zaujíma len výskyt javu a jeho absencia je nerozlíšiteľná od jeho nezmerania. Pritom

- maximálna hodnota agregačnej funkcie sa má dosiahnuť vtedy, keď počet hodnôt 1 je najväčší možný,
- minimálna hodnota agregáciu má nastať v prípade, keď sa v dátach nevyskytuje žiadna hodnota 1 – či už ide o samé 0, samé NaN, alebo ich ľubovoľnú kombináciu, čo odráža symetrické postavenie hodnôt 0 a NaN.

Navrhovaná agregačná funkcia má tvar:

$$A(\#0, \#1, \#\mathrm{NaN}) = \frac{\#1}{\#0 + \#1 + \#\mathrm{NaN}}$$

Z matematického hľadiska ide o limitný prípad predchádzajúcej funkcie pre $w \to 0^+$. V tomto modeli dochádza k totálnej strate rozlišovacej schopnosti medzi „známou nulou“ a „neznámou hodnotou“, čo zjednodušuje výpočet v binárnych klasifikátoroch s vysokou mierou neurčitosti. Teda, takto definovaná agregácia poskytuje jednoduchý a interpretačne stabilný spôsob hodnotenia dát v situáciách, keď je relevantná iba prítomnosť hodnoty 1 a všetky ostatné

hodnoty predstavujú rovnocennú formu nerelevantnosti alebo absencie pozitívnej informácie. Takáto agregačná funkcia zároveň spĺňa požadované vlastnosti monotónnosti (pri fixnom N):

- je rastúca vzhľadom na rastúci počet 1, keďže každá ďalšia jednotka zvyšuje podiel relevantnej informácie,
- je konštantná vzhľadom na rastúci počet 0 a NaN pri fixnom počte hodnôt 1, keďže tieto dve hodnoty sú symetrické, a teda ich vzájomné nahrádzanie nemení výsledný podiel hodnôt 1.

4.4 Zhrnutie a komparácia

Všetky tri uvedené funkcie možno zapísať v zovšeobecnenom tvare:

$$A(\#0, \#1, \#\mathrm{NaN}) = \frac{\alpha \cdot \#1 + \beta \cdot \#0 + \gamma \cdot \#\mathrm{NaN}}{\#0 + \#1 + \#\mathrm{NaN}}$$

kde pre jednotlivé prípady volíme koeficienty (α, β, γ) z množiny $\{0, w, 1\}$. Tento zjednotený pohľad umožňuje implementáciu jednej modulárnej funkcie v programových prostrediach pre dátovú vedu, kde sa správanie agregácie mení len nastavením váhových parametrov v závislosti od sémantiky problému.

5 Bibliografia

[1] Zimmerman, E. (n.d.). *Nástroje Erica Zimmermana*. Dostupné na https://ericzimmerman.github.io/#!index.md (prístupné 15. apríla 2026).

[2] Zimmerman, E. (n.d.). *EvtxECmd*. Repozitár GitHub. Dostupné na https://github.com/EricZimmerman/evtx (prístupné 15. apríla 2026).

[3] Zimmerman, E. (n.d.). *MFTECmd*. Repozitár GitHub. Dostupné na https://github.com/EricZimmerman/MFTECmd (prístupné 15. apríla 2026).

[4] Zimmerman, E. (n.d.). *RECmd*. Repozitár GitHub. Dostupné na https://github.com/EricZimmerman/RECmd (prístupné 15. apríla 2026).

[5] Zimmerman, E. (n.d.). *PECmd*. Repozitár GitHub. Dostupné na https://github.com/EricZimmerman/PECmd (prístupné 15. apríla 2026).

[6] Zimmerman, E. (n.d.). AppCompatCacheParser. Repozitár GitHub. Dostupné na https://github.com/ericzimmerman/appcompatcacheparser (prístupné 15. apríla 2026).

[7] Zimmerman, E. (n.d.). *JLECmd*. Repozitár GitHub. Dostupné na https://github.com/EricZimmerman/JLECmd (prístupné 15. apríla 2026).