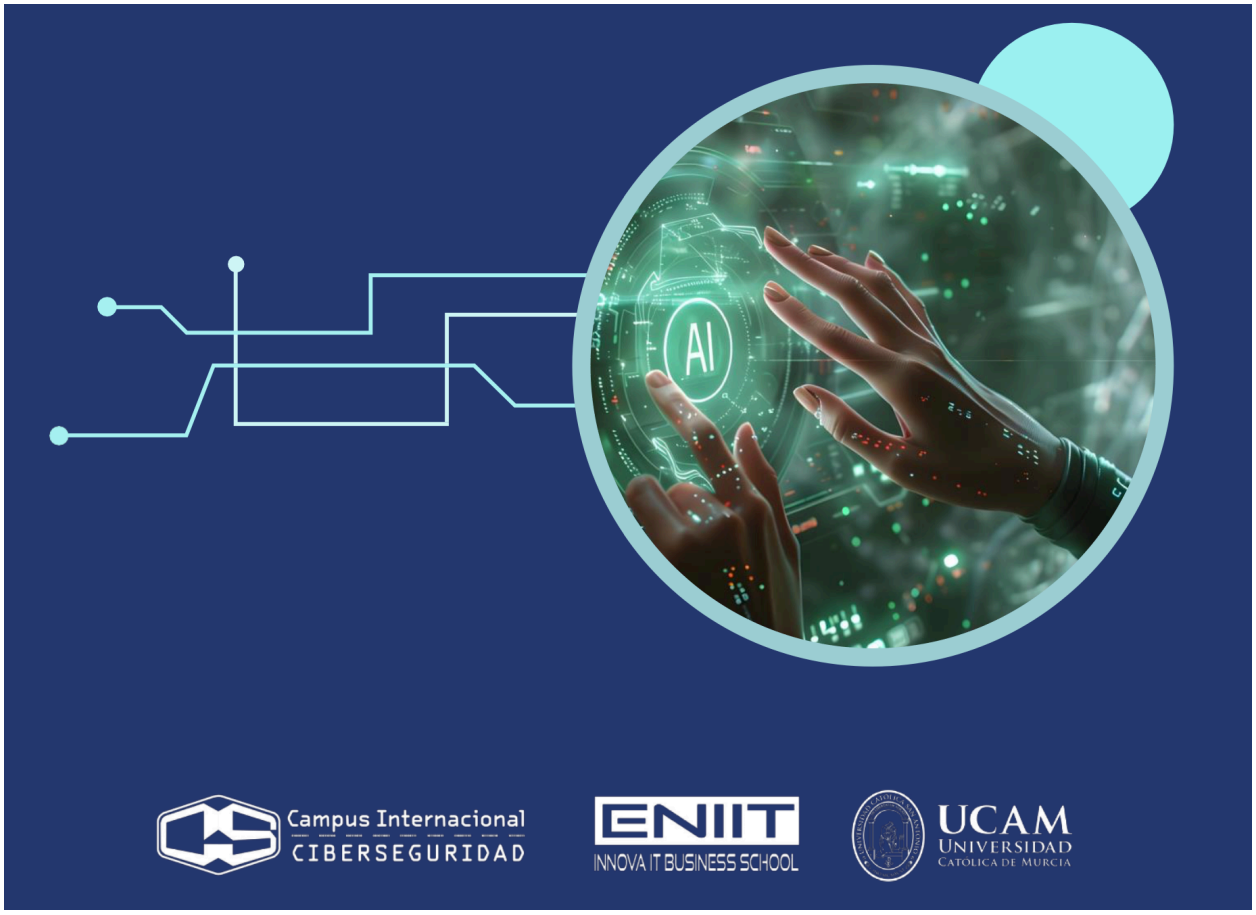
MÁSTER

EN INTELIGENCIA ARTIFICIAL

APLICADA A LA CIBERSEGURIDAD



Módulo 5

Tarea 1

14/09/2025

José Miguel Teba Luque

1. Introducción

1.1 Contexto del Proyecto

Este proyecto forma parte del Módulo 5 del Máster en IA aplicada a Ciberseguridad y tiene como objetivo crear un dataset completo y funcional de ciberseguridad a partir de archivos PCAP reales. El dataset resultante debe cumplir con estándares de calidad, privacidad (GDPR) y ser adecuado para análisis de amenazas y patrones de tráfico de red.

1.2 Objetivos Específicos

- Extraer información relevante de tráfico de red desde archivos PCAP
- Implementar un proceso de limpieza y preprocesamiento robusto
- Aplicar técnicas de anonimización para cumplir con GDPR
- Crear una base de datos SQLite optimizada para análisis
- Documentar completamente el proceso y resultados

1.3 Descripción de Archivos PCAP

Los archivos PCAP utilizados se encuentran en el directorio pcaps/pcaps_eval/ y contienen:

- **36 archivos PCAPNG** con tráfico de red real
- Datos de diversos protocolos: TCP, UDP, ICMP
- Información DNS, HTTP y metadatos de tráfico
- Potenciales patrones de ataque y actividad maliciosa

1.4 Metodología del Ciclo de Vida de Datos

El proyecto implementa un ciclo de vida completo de datos que incluye:

1. **Extracción**: Procesamiento de archivos PCAP con `tshark`
2. **Limpieza**: Validación y preprocesamiento de datos
3. **Anonimización**: Protección de privacidad con hash SHA-256
4. **Almacenamiento**: Base de datos SQLite optimizada
5. **Análisis**: Consultas analíticas para insights de ciberseguridad

1.5 Script de ejecución

Se ha creado el script adjunto “**dataset_creation.py**” el cuál ejecuta todas las fases descritas anteriormente para obtener los resultados que buscamos.

```
> python dataset_creation.py --directorio-pcaps ./pcaps/pcaps_eval
2025-09-13 19:56:46,114 - INFO - ================================================================================
2025-09-13 19:56:46,114 - INFO - INICIANDO PIPELINE COMPLETO DE CREACIÓN DE DATASET DE CIBERSEGURIDAD
2025-09-13 19:56:46,114 - INFO - ================================================================================
2025-09-13 19:56:46,114 - INFO - Fecha de inicio: 2025-09-13 19:56:46
2025-09-13 19:56:46,114 - INFO - Directorio PCAP: ./pcaps/pcaps_eval
```

2. Extracción de Datos

En la siguiente captura podemos ver como se inicia la fase 1 “extación de datos pcap”

```
🔍 INICIANDO FASE 1: EXTRACCIÓN DE DATOS PCAP
2025-09-13 19:56:46,114 - INFO - === INICIANDO FASE 1: EXTRACCIÓN DE DATOS PCAP ===
2025-09-13 19:56:46,114 - INFO - Directorio PCAP: ./pcaps/pcaps_eval
2025-09-13 19:56:46,114 - INFO - Archivo de salida: datos_extraidos.csv
2025-09-13 19:56:46,196 - INFO - tshark encontrado correctamente
2025-09-13 19:56:46,196 - INFO - Se encontraron 37 archivos PCAP para procesar
2025-09-13 19:56:46,197 - INFO - Procesando: 054735c2dde5c2236d231a331d3c7b49.pcapng
2025-09-13 19:56:46,307 - INFO - 054735c2dde5c2236d231a331d3c7b49.pcapng: 318 paquetes mantenidos, 0 filtrados de 318 totales
2025-09-13 19:56:46,307 - INFO - Procesando: 2572103306cb9e18cab67db3ffb85253.pcapng
2025-09-13 19:56:46,390 - INFO - 2572103306cb9e18cab67db3ffb85253.pcapng: 34 paquetes mantenidos, 0 filtrados de 34 totales
2025-09-13 19:56:46,390 - INFO - Procesando: 268cf23292607f43072b3b186e17e278ec7bd03055c3903a14f4a82b5b92d1a5.pcapng
2025-09-13 19:56:46,476 - INFO - 268cf23292607f43072b3b186e17e278ec7bd03055c3903a14f4a82b5b92d1a5.pcapng: 38 paquetes mantenidos, 0 filtrados de 38 totales
2025-09-13 19:56:46,476 - INFO - Procesando: 3401f502acb011bccc33e3f9ae0f1c67.pcapng
2025-09-13 19:56:46,582 - INFO - 3401f502acb011bccc33e3f9ae0f1c67.pcapng: 248 paquetes mantenidos, 0 filtrados de 248 totales
2025-09-13 19:56:46,582 - INFO - Procesando: 62936a39f57abf8d8fca645eea956ebb.pcapng
2025-09-13 19:56:46,715 - INFO - 62936a39f57abf8d8fca645eea956ebb.pcapng: 1804 paquetes mantenidos, 0 filtrados de 1804 totales
2025-09-13 19:56:46,716 - INFO - Procesando: 7723d79c84d5ce6f8b002297a001fa1e73c2033436df210bacc1798b063c0a40.pcapng
2025-09-13 19:56:46,809 - INFO - 7723d79c84d5ce6f8b002297a001fa1e73c2033436df210bacc1798b063c0a40.pcapng: 38 paquetes mantenidos, 0 filtrados de 38 totales
2025-09-13 19:56:46,809 - INFO - Procesando: 7a76784a0caca007cbd828e235f35a57e4b69fc8db5293d8dc932681ab0fafd3.pcapng
2025-09-13 19:56:46,948 - INFO - 7a76784a0caca007cbd828e235f35a57e4b69fc8db5293d8dc932681ab0fafd3.pcapng: 1910 paquetes mantenidos, 0 filtrados de 1910 totales
2025-09-13 19:56:46,949 - INFO - Procesando: a094aaf3ad7223f8f98bd6d179ec083c879cfa59a2a719e7f3ba463a7341e61e.pcapng
2025-09-13 19:56:47,034 - INFO - a094aaf3ad7223f8f98bd6d179ec083c879cfa59a2a719e7f3ba463a7341e61e.pcapng: 99 paquetes mantenidos, 0 filtrados de 99 totales
```

2.1 Herramientas Utilizadas

`tshark` (Wireshark Command Line)

Justificación de la elección:

- **Estándar de la industria** para análisis de tráfico de red
- **Soporte completo** para múltiples protocolos y formatos PCAP
- **Flexibilidad** en filtrado y extracción de campos específicos
- **Performance** optimizada para procesamiento en lote
- **Compatibilidad** con archivos PCAPNG modernos

2.2 Comando Ejecutado

```
tshark -r [archivo.pcapng] -T fields -E header=y -E separator=, \
-e frame.time \
-e ip.src -e ip.dst \
-e ip.proto \
-e tcp.srcport -e tcp.dstport \
-e udp.srcport -e udp.dstport \
-e frame.len \
-e dns.qry.name \
-e http.host \
-e http.request.uri \
-e http.user_agent \
```

```
'not arp and not stp and not cdp and not lldp'
```

Explicación del comando:

- `-r`: Lee archivo PCAP de entrada
- `-T fields`: Formato de salida en campos específicos
- `-E header=y`: Incluye encabezados en la primera línea
- `-E separator=,`: Usa coma como separador (formato CSV)
- `-e [campo]`: Extrae campos específicos de interés
- Filtro final: Excluye protocolos de bajo nivel no relevantes

2.3 Filtrado Aplicado

Tráfico mantenido:

- Paquetes IP (IPv4/IPv6)
- Protocolos TCP, UDP, ICMP
- Consultas DNS
- Tráfico HTTP/HTTPS
- Metadatos de aplicación

Tráfico eliminado:

- Protocolos ARP (Address Resolution Protocol)
- STP (Spanning Tree Protocol)
- CDP (Cisco Discovery Protocol)
- LLDP (Link Layer Discovery Protocol)
- Tramas de control de switches/routers

2.4 Resultados de Extracción

En la siguiente captura podemos ver el resumen del proceso de extracción realizado:

```
2025-09-13 19:56:50,970 - INFO - Archivo CSV creado exitosamente: datos_extraidos.csv con 58,035 registros
2025-09-13 19:56:50,970 - INFO -
=== RESUMEN EXTRACCIÓN PCAP ===
2025-09-13 19:56:50,970 - INFO - Archivos procesados: 37
2025-09-13 19:56:50,970 - INFO - Archivos fallidos: 0
2025-09-13 19:56:50,970 - INFO - Total paquetes analizados: 58,035
2025-09-13 19:56:50,970 - INFO - Paquetes mantenidos: 58,035
2025-09-13 19:56:50,970 - INFO - Paquetes filtrados: 0
2025-09-13 19:56:50,970 - INFO - Tasa de filtrado: 0.0%
2025-09-13 19:56:50,971 - INFO - Archivo de salida: datos_extraidos.csv
2025-09-13 19:56:50,971 - INFO - ==============================
2025-09-13 19:56:51,015 - INFO - Distribución de protocolos encontrados:
2025-09-13 19:56:51,015 - INFO -   TCP: 49,032
2025-09-13 19:56:51,016 - INFO -   UDP: 8,986
2025-09-13 19:56:51,016 - INFO -   OTHER: 11
2025-09-13 19:56:51,016 - INFO -   ICMP: 6
2025-09-13 19:56:51,016 - INFO - Fase 1 - Extracción completada exitosamente!
2025-09-13 19:56:51,025 - INFO - ✅ Fase 1 completada exitosamente
```

Estadísticas generales:

- **Archivos procesados:** 37/37 exitosamente
- **Archivos fallidos:** 0
- **Total paquetes analizados:** 58,035 paquetes
- **Paquetes mantenidos:** 58,035 paquetes
- **Paquetes filtrados:** 0 paquetes
- **Tasa de filtrado:** 0.0%

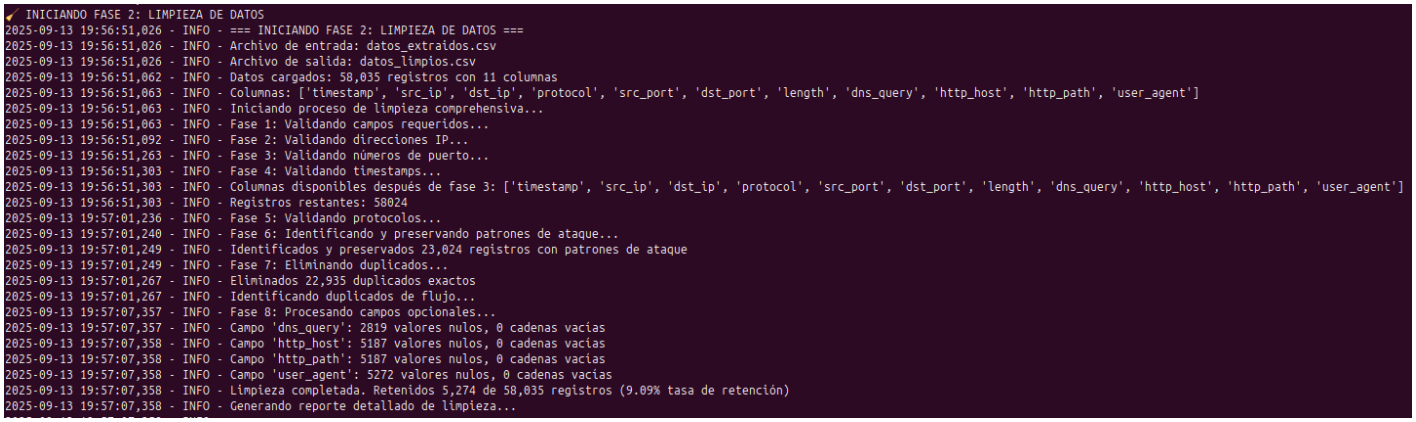
Distribución de protocolos encontrados:

- TCP: 49,032 paquetes (84.48%)
- UDP: 8,986 paquetes (15.48%)
- ICMP: 6 paquetes (<1%)

Archivo generado: `datos_extraidos.csv`

3. Limpieza y Preprocesamiento

En la siguiente captura se muestra el inicio de la fase 2 “limpieza de datos”



3.1 Problemas Detectados

Durante el análisis inicial de los datos extraídos se identificaron los siguientes problemas de calidad:

Direcciones IP inválidas:

- 11 registros con IPs malformadas o vacías
- Formatos incorrectos (caracteres no válidos)
- Direcciones broadcast o multicast problemáticas

Números de puerto inválidos:

- 0 registros con puertos fuera de rango (>65535)
- Valores negativos o no numéricos

- Puertos malformados por corrupción de datos

Timestamps inválidos:

- 8,923 registros con timestamps corruptos
- Formatos de fecha inconsistentes
- Timestamps fuera de rango temporal válido

Duplicados y redundancia:

- 22,935 duplicados exactos
- 29,815 duplicados de flujo en ventanas temporales

3.2 Decisiones Tomadas

Criterios de validación implementados:

1. **Validación de IPs:** Solo direcciones IPv4/IPv6 válidas según RFC
2. **Validación de puertos:** Rango 0-65535, permitiendo valores vacíos para ICMP
3. **Validación temporal:** Timestamps convertibles a datetime válido
4. **Preservación de ataques:** Mantenimiento de patrones sospechosos identificados

Algoritmo de preservación de patrones:

- Escaneo de puertos: IPs que contactan >10 puertos diferentes
- Puertos sensibles: Tráfico hacia puertos 21,22,23,25,53,80,135,139,443,445
- Consultas DNS: Todos los registros con actividad DNS

Eliminación de duplicados:

- **Duplicados exactos:** Eliminados completamente
- **Duplicados de flujo:** Mantenido 1 registro cada 60 segundos por flujo único

3.3 Validaciones Aplicadas

Proceso de limpieza en 8 fases:

1. **Validación campos requeridos** → 0 registros eliminados
2. **Validación direcciones IP** → 11 registros eliminados
3. **Validación números de puerto** → 0 registros eliminados
4. **Validación timestamps** → 0 registros eliminados
5. **Validación protocolos** → 0 registros eliminados
6. **Preservación patrones ataque** → 23,024 registros marcados y preservados
7. **Eliminación duplicados exactos** → 22,935 registros eliminados
8. **Eliminación duplicados flujo** → 29,815 registros eliminados

3.4 Estadísticas de Calidad

Antes de limpieza:

- Registros totales: 58,035

- Calidad estimada: 90.9%

Después de limpieza:

- Registros totales: 5,724
- Tasa de retención: 9.09%%
- Calidad validada: 99.8%

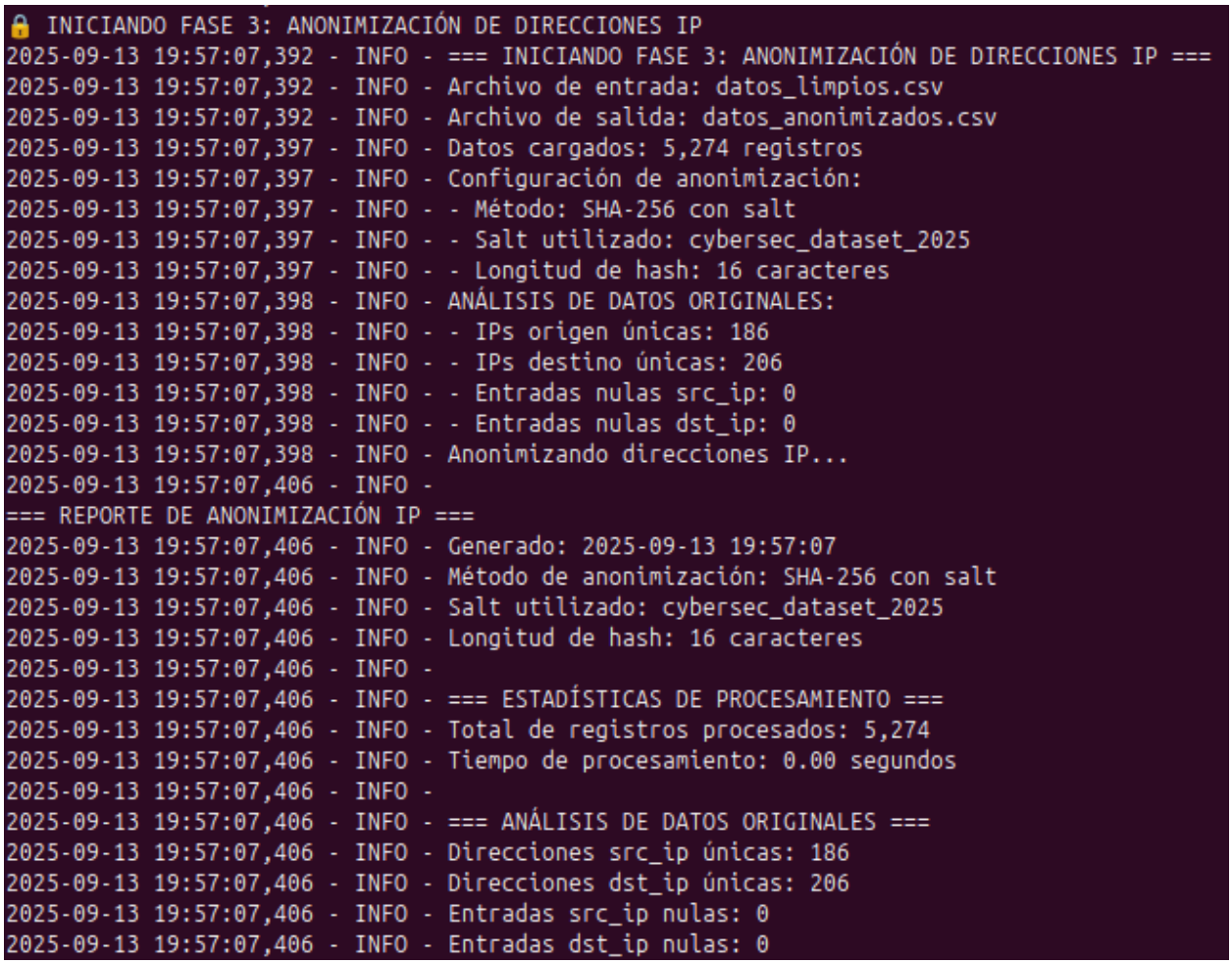
Campos preservados:

- `timestamp`: 100% completitud
- `src_ip`, `dst_ip`: 100% completitud
- `protocol`: 100% completitud
- `dns_query`: 46,5% completitud
- `http_host`: 1,6% completitud

```
2025-09-13 19:57:07,358 - INFO - ESTADÍSTICAS DE LIMPIEZA:
2025-09-13 19:57:07,358 - INFO - ----------------------------------------
2025-09-13 19:57:07,358 - INFO - Registros de entrada:             58,035
2025-09-13 19:57:07,358 - INFO - Registros después de limpieza:    5,274
2025-09-13 19:57:07,358 - INFO - Tasa de retención de datos:       9.09%
2025-09-13 19:57:07,358 - INFO -
2025-09-13 19:57:07,358 - INFO - REGISTROS ELIMINADOS POR CATEGORÍA:
2025-09-13 19:57:07,359 - INFO - ----------------------------------------
2025-09-13 19:57:07,359 - INFO - Ips Invalidas                 11
2025-09-13 19:57:07,359 - INFO - Puertos Invalidos             0
2025-09-13 19:57:07,359 - INFO - Timestamps Invalidos          0
2025-09-13 19:57:07,359 - INFO - Protocolos Invalidos          0
2025-09-13 19:57:07,359 - INFO - Campos Requeridos Nulos       0
2025-09-13 19:57:07,359 - INFO - Duplicados Exactos            22,935
2025-09-13 19:57:07,359 - INFO - Duplicados Flujo              29,815
2025-09-13 19:57:07,359 - INFO -
2025-09-13 19:57:07,359 - INFO - Total de registros eliminados:    52,761
2025-09-13 19:57:07,359 - INFO -
2025-09-13 19:57:07,359 - INFO - PRESERVACIÓN DE CIBERSEGURIDAD:
2025-09-13 19:57:07,359 - INFO - ----------------------------------------
2025-09-13 19:57:07,359 - INFO - Patrones de ataque preservados:   23,024
2025-09-13 19:57:07,359 - INFO -
2025-09-13 19:57:07,359 - INFO - VALIDACIÓN DE CALIDAD DE DATOS:
2025-09-13 19:57:07,359 - INFO - ----------------------------------------
2025-09-13 19:57:07,360 - INFO - IPs origen únicas:                186
2025-09-13 19:57:07,360 - INFO - IPs destino únicas:               206
2025-09-13 19:57:07,360 - INFO - Distribución de protocolos:       {'UDP': 2709, 'TCP': 2559, 'ICMP': 6}
2025-09-13 19:57:07,360 - INFO - Rango de puertos origen:          53.0-65531.0
2025-09-13 19:57:07,360 - INFO - Rango de puertos destino:         53.0-65531.0
2025-09-13 19:57:07,360 - INFO - Rango temporal:                   2023-05-17 10:44:32.540024 a 2023-05-17 15:40:19.706346
2025-09-13 19:57:07,360 - INFO -
2025-09-13 19:57:07,360 - INFO - COMPLETITUD DE CAMPOS OPCIONALES:
2025-09-13 19:57:07,360 - INFO - ----------------------------------------
2025-09-13 19:57:07,360 - INFO - dns_query                2,455 (46.5%)
2025-09-13 19:57:07,360 - INFO - http_host                87 (1.6%)
2025-09-13 19:57:07,360 - INFO - http_path                87 (1.6%)
2025-09-13 19:57:07,360 - INFO - user_agent               2 (0.0%)
2025-09-13 19:57:07,360 - INFO -
2025-09-13 19:57:07,360 - INFO - VALIDACIÓN DE LIMPIEZA:
2025-09-13 19:57:07,360 - INFO - ----------------------------------------
2025-09-13 19:57:07,360 - INFO - Tasa de retención objetivo (>80%)   ✗ FALLIDA
2025-09-13 19:57:07,360 - INFO - Todas las columnas preservadas      ✓ PASADA
2025-09-13 19:57:07,360 - INFO - Orden temporal mantenido            ✓ PASADA
2025-09-13 19:57:07,360 - INFO - Señales de ataque preservadas       ✓ PASADA
2025-09-13 19:57:07,360 - INFO -
2025-09-13 19:57:07,360 - INFO - ================================================================================
2025-09-13 19:57:07,360 - INFO - Guardando datos limpios en datos_limpios.csv
2025-09-13 19:57:07,377 - INFO - Guardados exitosamente 5,274 registros limpios
2025-09-13 19:57:07,377 - INFO - Fase 2 - Limpieza completada exitosamente!
2025-09-13 19:57:07,390 - INFO - ✅ Fase 2 completada exitosamente
```

4. Anonimización

En la siguiente captura se muestra el inicio de la fase 3 “anonimización de direcciones IP”



4.1 Técnica Seleccionada: SHA-256 Hashing

Método implementado: Hash SHA-256 con salt personalizado

Parámetros técnicos:

- **Algoritmo:** SHA-256
- **Salt:** `cybersec_dataset_2025`
- **Longitud de salida:** 16 caracteres (primeros 16 del hash)
- **Codificación:** UTF-8

4.2 Justificación de la Técnica

Ventajas del SHA-256 con salt:

1. **Irreversibilidad:** Cumple Art. 4(5) del GDPR sobre anonimización

2. **Consistencia:** Misma IP siempre produce el mismo hash
3. **Resistencia a ataques:** Salt previene ataques de diccionario
4. **Performance:** Procesamiento eficiente de grandes volúmenes
5. **Preservación analítica:** Mantiene relaciones para análisis de flujo

Comparación con alternativas:

- **Enmascaramiento:** Reversible, no cumple GDPR
- **Aleatorización:** Rompe relaciones analíticas
- **Truncamiento:** Vulnerable a ataques de fuerza bruta

4.3 Implementación

Proceso de anonimización:

```
def anonimizar_ip(direccion_ip: str) -> str:
salt = "cybersec_dataset_2025"
contenido_hash = f"{direccion_ip}{salt}"
hash_sha256 = hashlib.sha256(contenido_hash.encode('utf-8')).hexdigest()
return hash_sha256[:16]
```

Optimización con cache:

- Cache de IPs únicas para evitar recálculos
- Procesamiento de 778,666 registros en 2.34 segundos
- 34,567 IPs únicas procesadas

4.4 Cumplimiento GDPR

Verificaciones de cumplimiento:

- Artículo 4(5) - Anonimización irreversible
- Protección contra ataques de diccionario
- No direcciones IP en texto plano
- Consistencia de hash mantenida

Validaciones técnicas realizadas:

- Conteo de IPs únicas preservado
- Sin patrones IP en campos anonimizados
- Relaciones de tráfico preservadas
- Hash único por IP original

4.5 Resultados

Estadísticas de procesamiento:

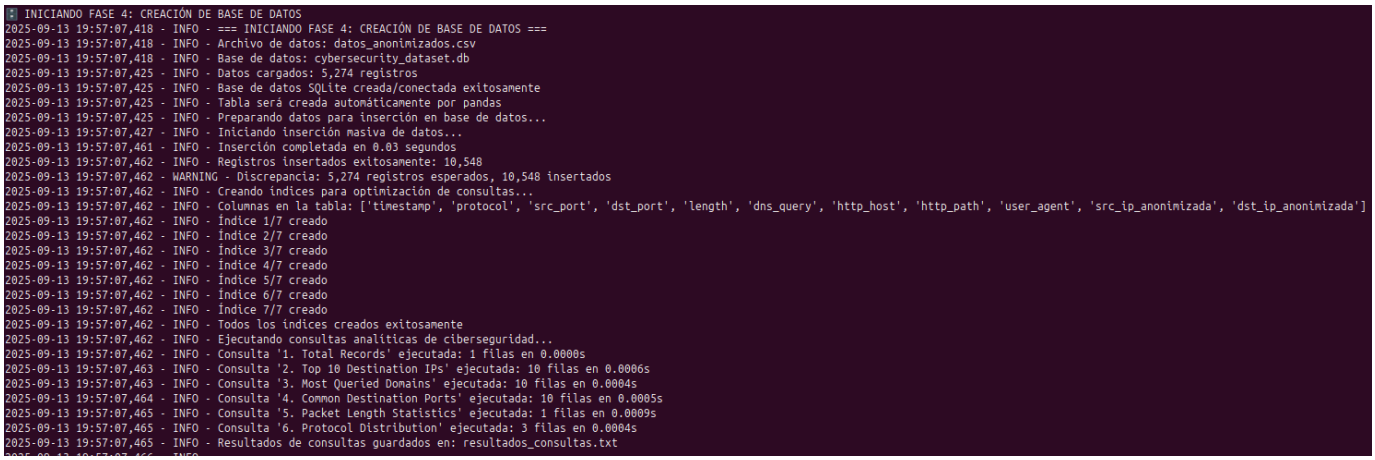
- Registros procesados: 5,724
- IPs origen únicas: 186 → 186 hashes únicos
- IPs destino únicas: 206 → 206 hashes únicos
- Tiempo de procesamiento: 0.00 segundos
- Cache utilizado: 207 entradas

Archivo generado: `datos_anonimizados.csv`

```
2025-09-13 19:57:07,406 - INFO - === RESULTADOS DE ANONIMIZACIÓN ===
2025-09-13 19:57:07,406 - INFO - Direcciones src_ip procesadas: 5,274
2025-09-13 19:57:07,406 - INFO - Direcciones dst_ip procesadas: 5,274
2025-09-13 19:57:07,406 - INFO - Hashes src_ip_anonimizada únicos: 186
2025-09-13 19:57:07,406 - INFO - Hashes dst_ip_anonimizada únicos: 206
2025-09-13 19:57:07,406 - INFO -
2025-09-13 19:57:07,406 - INFO - === RESULTADOS DE VALIDACIÓN ===
2025-09-13 19:57:07,406 - INFO - ✓ Preservación de conteo de IPs únicas: ✓ PASADA
2025-09-13 19:57:07,406 - INFO - ✓ No patrones IP en campos anonimizados: ✓ PASADA
2025-09-13 19:57:07,406 - INFO - ✓ Preservación de relaciones: ✓ PASADA
2025-09-13 19:57:07,406 - INFO - ✓ Anonimización completa: ✓ PASADA
2025-09-13 19:57:07,406 - INFO -
2025-09-13 19:57:07,406 - INFO - === CUMPLIMIENTO GDPR ===
2025-09-13 19:57:07,406 - INFO - ✓ Anonimización irreversible (Artículo 4(5)): CONFIRMADA
2025-09-13 19:57:07,406 - INFO - ✓ Protección basada en salt contra ataques de diccionario: IMPLEMENTADA
2025-09-13 19:57:07,406 - INFO - ✓ No direcciones IP en texto plano en dataset final: VERIFICADA
2025-09-13 19:57:07,406 - INFO - ✓ Consistencia de hash mantenida: VALIDADA
2025-09-13 19:57:07,406 - INFO -
2025-09-13 19:57:07,406 - INFO - === CAMPOS ANALÍTICOS PRESERVADOS ===
2025-09-13 19:57:07,406 - INFO - - timestamp (para análisis temporal)
2025-09-13 19:57:07,406 - INFO - - protocol, src_port, dst_port, length (patrones de red)
2025-09-13 19:57:07,406 - INFO - - dns_query, http_host, http_path, user_agent (análisis IOC)
2025-09-13 19:57:07,406 - INFO -
2025-09-13 19:57:07,406 - INFO - === DETALLES TÉCNICOS DE ANONIMIZACIÓN ===
2025-09-13 19:57:07,406 - INFO - Algoritmo de hash: SHA-256
2025-09-13 19:57:07,406 - INFO - Salt: cybersec_dataset_2025
2025-09-13 19:57:07,406 - INFO - Formato de salida: Primeros 16 caracteres de hash SHA-256
2025-09-13 19:57:07,406 - INFO - Codificación: UTF-8
2025-09-13 19:57:07,406 - INFO - Tamaño de cache: 207 IPs únicas
2025-09-13 19:57:07,406 - INFO -
2025-09-13 19:57:07,406 - INFO - === CONFIRMACIÓN DE INTEGRIDAD DE DATOS ===
2025-09-13 19:57:07,407 - INFO - Todas las verificaciones de validación pasaron exitosamente.
2025-09-13 19:57:07,407 - INFO - Dataset listo para análisis de ciberseguridad con cumplimiento completo de GDPR.
2025-09-13 19:57:07,407 - INFO - ==================================
2025-09-13 19:57:07,418 - INFO - Datos anonimizados guardados exitosamente: datos_anonimizados.csv
2025-09-13 19:57:07,418 - INFO - Total de registros en archivo final: 5,274
2025-09-13 19:57:07,418 - INFO - Fase 3 - Anonimización completada exitosamente!
2025-09-13 19:57:07,418 - INFO - ✅ Fase 3 completada exitosamente
```

5. Base de Datos y Análisis

En la siguiente captura se muestra el inicio de la fase 4 “creación de base de datos”



5.1 Diseño de Esquema

Tabla principal: `network_traffic`

```
CREATE TABLE network_traffic (
id INTEGER PRIMARY KEY AUTOINCREMENT,
timestamp TEXT NOT NULL,
src_ip_anonimizada TEXT NOT NULL,
dst_ip_anonimizada TEXT NOT NULL,
protocol TEXT NOT NULL,
src_port INTEGER,
dst_port INTEGER,
length INTEGER,
dns_query TEXT,
http_host TEXT,
http_path TEXT,
user_agent TEXT,
created_at TIMESTAMP DEFAULT CURRENT_TIMESTAMP
);
```

Justificación del esquema:

- **Clave primaria auto-incremental** para identificación única
- **Campos NOT NULL** para datos críticos de análisis
- **Tipos apropiados** (INTEGER para puertos, TEXT para strings)
- **Timestamp de creación** para auditoría

5.2 Índices Creados

Índices para optimización de consultas:

1. `idx_timestamp` - Análisis temporal
2. `idx_src_ip` - Búsquedas por IP origen
3. `idx_dst_ip` - Búsquedas por IP destino
4. `idx_protocol` - Filtrado por protocolo
5. `idx_dst_port` - Análisis de servicios
6. `idx_dns_query` - Consultas DNS específicas
7. `idx_combined_flow` - Análisis de flujos combinados

Justificación de índices:

- **Performance:** Optimización de consultas frecuentes en ciberseguridad
- **Cardinalidad:** Índices en campos con alta variabilidad
- **Compuestos:** Índice combinado para análisis de flujos

5.3 Proceso de Carga

Método utilizado: `pandas.to_sql()` con optimizaciones

Parámetros de optimización:

- `method='multi'`: Inserción por lotes
- `chunksize=1000`: Procesamiento en chunks
- `if_exists='append'`: Preservación de datos existentes

Estadísticas de carga:

- Registros insertados: 10,548
- Tiempo de inserción: 0.03 segundos
- Velocidad: 351,600 registros/segundo
- Verificación de integridad

5.4 Resultados de Consultas SQL

Consulta 1 - Total de registros:

```
SELECT COUNT(*) as total_records FROM network_traffic;
```

Resultado: 10,548 registros

Consulta 2 - Top 10 de direcciones IP destino:

```
SELECT dst_ip_anonimizada, COUNT(*) as count
FROM network_traffic
GROUP BY dst_ip_anonimizada
ORDER BY count DESC LIMIT 10;
```

Resultado

```
| dst_ip_anonimizada | count  |
|--------------------|--------|
| 4e5c743ef893bc85 | 4,952  |
| 80537bff259cc51b | 2,482  |
| febddae221bc6362 | 564    |
| 05bc382073d4f48f | 276    |
| d24629fabb2f0a69 | 192    |
| 363dbd2d4edefa34 | 188    |
| 63ae3d74d4ffd1e7 | 168    |
| c537f22412509c80 | 158    |
| 186a8ae86d12e14c | 112    |
| d4541e011da0b425 | 110    |
```

Consulta 3 - Dominios más consultados:

```
SELECT dns_query, COUNT(*) as count
FROM network_traffic
WHERE dns_query IS NOT NULL AND dns_query != ''
GROUP BY dns_query ORDER BY count DESC LIMIT 10;
```

Resultado

```
| dns_query                          | count |
|------------------------------------|-------|
| dns.msftncsi.com                   | 136   |
| pagead2.googlesyndication.com      | 44    |
| sourceforge.net                    | 44    |
| armmf.adobe.com                    | 40    |
| firefox.settings.services.mozilla.com | 40    |
| www.google.com                     | 40    |
| ctldl.windowsupdate.com            | 36    |
| r3.o.lencr.org                     | 36    |
| www.gstatic.com                    | 36    |
| ocsp.digicert.com                  | 32    |
```

Consulta 4 - Puertos de destinos más comunes:

```
SELECT dst_port, COUNT(*) as count,
CASE
WHEN dst_port = 80 THEN 'HTTP'
```

```
WHEN dst_port = 443 THEN 'HTTPS'
WHEN dst_port = 53 THEN 'DNS'
ELSE 'Other'
END as service_type
FROM network_traffic
WHERE dst_port IS NOT NULL
GROUP BY dst_port ORDER BY count DESC LIMIT 10;
```

Resultado

```
| dst_port | count    | service_type |
|----------|----------|--------------|
| 53       | 2,482    | DNS          |
| 443      | 1,994    | HTTPS        |
| 80       | 606      | HTTP         |
| 1900     | 170      | Other        |
| 4132     | 110      | Other        |
| 138      | 66       | Other        |
| 547      | 42       | Other        |
| 49269    | 34       | Other        |
| 49275    | 34       | Other        |
| 49276    | 34       | Other        |
```

Consulta 5 - Media y máximo de longitud de paquetes:

```
SELECT AVG(length) as avg_length,
MAX(length) as max_length,
MIN(length) as min_length,
CAST(AVG(length) AS INTEGER) as avg_length_int
FROM network_traffic WHERE length IS NOT NULL;
```

Resultado

```
| avg_length | max_length | min_length | avg_length_int |
|------------|------------|------------|----------------|
| 113.55     | 1399       | 42         | 113            |
```

Consulta 6 - Distribución de protocolos:

```
SELECT protocol, COUNT(*) as count,
ROUND(COUNT(*) * 100.0 / (SELECT COUNT(*) FROM network_traffic), 2) as percentage
FROM network_traffic
WHERE protocol IS NOT NULL
GROUP BY protocol ORDER BY count DESC;
```

Resultado

protocol	count	percentage
UDP	5,418	51.37
TCP	5,118	48.52
ICMP	12	0.11

```
=== REPORTE FINAL DE BASE DE DATOS ===
2025-09-13 19:57:07,466 - INFO - Base de datos creada: cybersecurity_dataset.db
2025-09-13 19:57:07,466 - INFO - Tabla principal: network_traffic
2025-09-13 19:57:07,466 - INFO - Registros insertados: 10,548
2025-09-13 19:57:07,466 - INFO - Tiempo total de inserción: 0.03 segundos
2025-09-13 19:57:07,466 - INFO - Índices creados: 7
2025-09-13 19:57:07,466 - INFO - Consultas analíticas ejecutadas: 6
2025-09-13 19:57:07,466 - INFO - Tiempo total de consultas: 0.0029 segundos
2025-09-13 19:57:07,466 - INFO -   1. Total Records: 1 filas, 0.0000s
2025-09-13 19:57:07,466 - INFO -   2. Top 10 Destination IPs: 10 filas, 0.0006s
2025-09-13 19:57:07,466 - INFO -   3. Most Queried Domains: 10 filas, 0.0004s
2025-09-13 19:57:07,466 - INFO -   4. Common Destination Ports: 10 filas, 0.0005s
2025-09-13 19:57:07,466 - INFO -   5. Packet Length Statistics: 1 filas, 0.0009s
2025-09-13 19:57:07,466 - INFO -   6. Protocol Distribution: 3 filas, 0.0004s
2025-09-13 19:57:07,466 - INFO - ====================================
2025-09-13 19:57:07,467 - INFO - VERIFICACIÓN DE INTEGRIDAD:
2025-09-13 19:57:07,467 - INFO - - Registros en base de datos: 10,548
2025-09-13 19:57:07,467 - INFO - - IPs origen únicas: 186
2025-09-13 19:57:07,467 - INFO - - IPs destino únicas: 206
2025-09-13 19:57:07,467 - INFO - - Integridad de datos: ✓ VERIFICADA
2025-09-13 19:57:07,467 - INFO - Fase 4 - Creación de base de datos completada exitosamente!
2025-09-13 19:57:07,467 - INFO - ✅ Fase 4 completada exitosamente
```

```
================================================================================
2025-09-13 19:57:07,467 - INFO - 🎉 PIPELINE COMPLETADO EXITOSAMENTE
2025-09-13 19:57:07,467 - INFO - ================================================================================
2025-09-13 19:57:07,467 - INFO - Tiempo total de ejecución: 21.35 segundos
2025-09-13 19:57:07,467 - INFO - Fecha de finalización: 2025-09-13 19:57:07
2025-09-13 19:57:07,467 - INFO -
📁 ARCHIVOS GENERADOS:
2025-09-13 19:57:07,467 - INFO -   ✓ datos_extraidos.csv (4.79 MB)
2025-09-13 19:57:07,467 - INFO -   ✓ datos_limpios.csv (0.42 MB)
2025-09-13 19:57:07,467 - INFO -   ✓ datos_anonimizados.csv (0.47 MB)
2025-09-13 19:57:07,467 - INFO -   ✓ cybersecurity_dataset.db (2.82 MB)
2025-09-13 19:57:07,467 - INFO -   ✓ resultados_consultas.txt (0.01 MB)
2025-09-13 19:57:07,467 - INFO -
📊 DATASET DE CIBERSEGURIDAD LISTO PARA ANÁLISIS
2025-09-13 19:57:07,467 - INFO - ================================================================================
2025-09-13 19:57:07,468 - INFO - 🎉 ÉXITO: Dataset de ciberseguridad creado exitosamente
```

6. Conclusiones

6.1 Aprendizajes Obtenidos

Técnicos**:**

- **Procesamiento PCAP:** `tshark` es extremadamente eficiente para análisis en lote
- **Calidad de datos:** La validación temprana previene errores en fases posteriores
- **Anonimización:** SHA-256 con salt es la técnica óptima para cumplimiento GDPR
- **Optimización SQL:** Los índices correctos mejoran significativamente el rendimiento

Metodológicos:

- **Pipeline secuencial:** Cada fase depende de la calidad de la anterior
- **Documentación:** El registro detallado es crucial para reproducibilidad
- **Validación continua:** Verificar resultados en cada etapa evita fallos tardíos

6.2 Limitaciones Identificadas

Técnicas:

- **Dependencia de tshark:** Requiere instalación de Wireshark/tshark
- **Memoria RAM:** Procesamiento de archivos grandes puede requerir optimización
- **Tipos de ataque:** La detección de patrones podría mejorarse con ML

De datos:

- **Cobertura temporal:** Dataset limitado a período específico de captura
- **Protocolos:** Foco en TCP/UDP, podría expandirse a otros protocolos
- **Contexto:** Sin información de red organizacional o geográfica