Imagen que contiene dibujo

Descripción generada automáticamente**Icono

Descripción generada automáticamente**

*JAVIER CUARTERO CORREDOR & GONZALO DE LOS REYES SÁNCHEZ*

*BASES DE DATOS AVANZADAS | ESCUELA SUPERIOR DE INFORMATICA, UCLM*

*TRABAJO 1*

DATA WAREHOUSE | ATAQUES DE CIBERSEGURIDAD

Tabla de contenido

[1. Resumen 2](#_Toc191240696)

[2. Introducción 2](#_Toc191240697)

[3. Desarrollo 3](#_Toc191240698)

[3.1 Procesamiento de datos 3](#_Toc191240699)

[3.2 Explicación tecnológica 4](#_Toc191240700)

[3.3 Diseño en estrella del Data Warehouse 4](#_Toc191240701)

[4. Resultados 5](#_Toc191240702)

[4.1 Creación y estructuración del Data Warehouse 5](#_Toc191240703)

[4.2 Procesamiento de datos exitoso. 6](#_Toc191240704)

[4.3 Inserción de datos exitosa 8](#_Toc191240705)

[5. Usos y utilización 9](#_Toc191240706)

# Resumen

En esta primera entrega de la asignatura se tiene como objetivo general crear una serie de programas/herramientas que permitan manejar los datos en su formato original y poder ejecutar Extracción, Transformación y Carga de datos en el Data Warehouse. Además, se propondrá un diseño en estrella con sus tablas de dimensiones y de hechos para posteriormente poder mostrar una serie de datos extraídos tal y como se ha diseñado.

# Introducción

El objetivo de este trabajo, como bien se ha comentado en el apartado anterior, es crear una serie de programas/herramientas que permitan el manejo de datos relacionados con ciberataques, implementado un proceso de ETL (Extracción, Transformación y Carga) en un Data Warehouse.

Para ello, se ha trabajado con un dataset en formato CSV denominado “Cyber Security Attacks”, el cual contiene información sobre 40000 ciberataques. Este conjunto de datos cuenta con 25 columnas, cada una representando distintas características:

* **Timestamp**: El tiempo en el que la actividad de red ocurrió.
* **Source** **IP** **Address**: La IP del emisor o iniciador del tráfico de red.
* **Destination** **IP** **Address**: La dirección IP del receptor o destino del tráfico de red.
* **Source** **Port**: El número de puerto utilizado por la dirección IP de origen.
* **Destination** **Port**: El número de puerto utilizado por la dirección IP de destino.
* **Protocol**: El protocolo de red usado (TCP, UDP, ICMP).
* **Packet** **Length**: Tamaño en bytes del paquete.
* **Packet** **Type**: Tipo de paquete (data packet, control packet).
* **Traffic** **Type**: Tipo de tráfico (web traffic, email traffic).
* **Payload** **Data**: Los datos transmitidos en el paquete.
* **Malware** **Indicators**: Indicadores de actividad potencialmente maliciosa o presencia de malware.
* **Anomaly** **Scores**: Puntuaciones que indican desviaciones del comportamiento esperado, utilizadas para la detección de anomalías.
* **Alerts**/**Warnings**: Notificaciones o avisos generados por sistemas de seguridad o herramientas de monitorización.
* **Attack** **Type**: Tipo de ataque detectado o supuesto (DDoS, SQL injection).
* **Attack** **Signature**: Patrones o firmas específicas asociadas con ataques conocidos.
* **Action** **Taken**: Acciones realizadas en respuesta a amenazas o anomalías detectadas.
* **Severity Level**: El nivel de gravedad asociado con una alerta o evento (por ejemplo, bajo, medio, alto).
* **User Information**: Información sobre el usuario involucrado en la actividad de la red.
* **Device Information**: Información sobre el dispositivo involucrado en la actividad de la red (por ejemplo, tipo de dispositivo, sistema operativo).
* **Network Segment**: El segmento o subred de la red donde ocurrió la actividad.
* **Geo-location Data**: Información de ubicación geográfica asociada a direcciones IP.
* **Proxy Information**: Información sobre los servidores proxy involucrados en la comunicación de red.
* **Firewall Logs**: Registros generados por dispositivos de firewall que indican tráfico permitido o bloqueado.
* **IDS/IPS Alerts**: Alertas generadas por sistemas de detección de intrusiones (IDS) o sistemas de prevención de intrusiones (IPS) que indican actividad sospechosa o maliciosa.
* **Log Source**: La fuente u origen de la entrada del registro (por ejemplo, nombre del sistema o dispositivo de registro).

El objetivo final es construir un Data Warehouse que facilite la consulta y el análisis de estos datos, además de permitirnos responder algunas preguntas claves cómo:

* **¿Cuáles son los segmentos de red más atacados?**
* **¿Qué protocolos de comunicación son más atacados?**

Que responderemos al final de este documento, en el apartado .5.

Para lograrlo, se ha seguido un enfoque estructurado basado en lo visto en las clases de la asignatura.

# Desarrollo

## Procesamiento de datos

El primer paso en la creación del Data Warehouse ha sido el procesamiento de los datos, asegurando que estén limpios y estructurados antes de cargarlos en la base de datos.

Pasos realizados:

1. *Carga del dataset*

* Se ha importado el archivo CSV en Pandas para su análisis inicial.

1. *Limpieza de los datos*

* Identificación y eliminación de los valores nulos o inconsistentes.

Se define como valor inconsistente a cualquier dato dentro del dataset que no sigue el formato esperado, presenta contradicciones o es incoherente en el contexto del análisis de ciberataques.

* Reemplazo de los datos faltantes con valores por defecto.
* Conversión de los tipos de datos (fechas, números, cadenas).

1. *Transformaciones aplicadas*

- Normalización de datos para mantener coherencia en los valores.

- Creación de columnas derivadas como hora, día, mes y año a partir del timestamp.

- Clasificación de direcciones en privadas/públicas.

- Extracción de información de dispositivos y navegadores desde los logs.

1. *Generación del dataset final*

* El dataset limpio y transformado se guardó como “*cybersecurity\_attacks\_cleaned.csv*”, listo para ser cargado en la base de datos.

## Explicación tecnológica

Para llevar a cabo este proyecto se han utilizado diferentes herramientas y tecnologías que han permitido la implementación del ETL y la gestión del Data Warehouse.

Tecnologías:

* Python: para preprocesamiento y transformación de datos.
* Pandas: para manipulación y limpieza del dataset.
* MySQL: para creación del Data Warehouse y almacenamiento de datos.
* MySQL Connector: para conexión entre Python y MySQL para la carga de datos.
* Matplot: para representación de consultas de manera gráfica.

**Proceso ETL**

Se ha seguido un flujo de Extracción, Transformación y Carga (ETL) en tres fases principales:

1. Extracción:
   * Lectura del CSV y carga de datos en un DataFrame de Pandas.
2. Transformación:
   * Procesamiento de datos (limpieza, normalización y creación de nuevas columnas).
   * Enriquecimiento de datos con nuevas métricas gracias a las nuevas columnas generadas.
3. Carga:
   * Creación de una base de datos en MySQL (ciberseguridad\_db).
   * Diseño de un esquema de almacenamiento, concretamente diseño en estrella.
   * Inserción de datos en las tablas del Data Warehouse.

## Diseño en estrella del Data Warehouse

Para estructuras los datos de la manera más eficiente posible, se ha optado por un diseño en estrella, compuesto por una tabla de hechos y varias tablas de dimensiones.

**Modelo dimensional**

* **Tabla de hechos:**
  + *hechos\_ataques***:** almacena la información central sobre los ataques.
* **Tablas de dimensiones:**
  + *dim\_origen*: información sobre la IP y puerto de origen.
  + dim\_destino: información sobre la IP y puerto de destino.
  + *dim\_protocolo*: tipo de protocolo utilizado en el ataque.
  + *dim\_tipo\_trafico*: categoría del tráfico involucrado.
  + *dim\_malware*: indicadores de malware asociados al ataque.
  + *dim\_anomalia*: niveles de anomalía detectados.
  + *dim\_severidad*: nivel de severidad del ataque.
  + *dim\_dispositivo*: dispositivo involucrado en el ataque.
  + *dim\_segmento*: segmento de red afectado.
  + *dim\_geo*: ubicación geográfica del ataque.

**Diagrama del Data Warehouse**

El diseño en estrella se puede visualizar en el siguiente diagrama

Interfaz de usuario gráfica

El contenido generado por IA puede ser incorrecto.

Se ha decidido la utilización de un diseño en estrella ya que este modelo nos permite:

* + Facilidad para las consultas rápidas debido a que las dimensiones están desnormalizadas, por lo que mejora el rendimiento.
  + Optimización del almacenamiento ya que este diseño evita la redundancia de datos sin necesidad de múltiples ´joins´ complejos.

# Resultados

Tras la implementación del proceso ETL y la construcción del Data Warehouse, se han obtenido los siguientes resultados.

## Creación y estructuración del Data Warehouse

Se ha diseñado e implementado un Data Warehouse con un modelo en estrella, donde la tabla de *hechos\_ataques* centraliza la información de los ciberataques y se relaciona con 10 dimensiones.

Se han definido correctamente las claves primarias y foráneas, asegurando la integridad referencial de los datos.

Se ha automatizado la creación de la base de datos y las tablas en MySQL, permitiendo su reutilización en futuros despliegues.

En la siguiente captura se muestra todas las tablas de la base de datos, con sus columnas y sus respectivos tipos de datos.

Imagen que contiene Interfaz de usuario gráfica

El contenido generado por IA puede ser incorrecto.

## 4.2 Procesamiento de datos exitoso.

Se han procesado 40000 registros de ciberataques desde un archivo CSV, limpiando y estructurando los datos para su inserción en la base de datos.

Se implementaron múltiples validaciones y transformaciones para asegurar la consistencia de los datos:

* Eliminación de valores nulos y reemplazo con valores adecuados.
* Corrección de valores inconsistentes, como direcciones IP inválidas, puertos fuera de rango y errores tipográficos en nombres de protocolos.
* Conversión de formatos para cumplir con los tipos de datos esperados en MySQL.

En las siguientes capturas, se muestra una comparación del “antes vs después” de la limpieza de los datos tras la ejecución de nuestro código.

Texto

El contenido generado por IA puede ser incorrecto.

Texto

El contenido generado por IA puede ser incorrecto.

## 4.3 Inserción de datos exitosa

Interfaz de usuario gráfica

El contenido generado por IA puede ser incorrecto.

Interfaz de usuario gráfica

El contenido generado por IA puede ser incorrecto.

# 5. Usos y utilización

Este apartado tiene como objetivo ver algunos de los principales casos de uso y preguntas clave que este DW puede responder mediante consultas SQL y visualización de los datos.

Debido a que estamos llegando al límite de extensión del documento, solo vamos a mostrar dos consultas, pero realmente se podrían resolver bastantes cuestiones del estilo.

Por ejemplo:

**¿Cuáles son los segmentos de red más atacados?**  
Permite identificar qué segmentos de la infraestructura han sido objetivo de más intentos de ataque, facilitando la asignación de recursos de seguridad a las zonas más vulnerables.

Texto

El contenido generado por IA puede ser incorrecto.

Resultado de la consulta

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

Texto

El contenido generado por IA puede ser incorrecto.

**¿Qué protocolos de comunicación son más atacados?**  
Permite identificar qué protocolos están siendo explotados con más frecuencia y ajustar las configuraciones de seguridad en consecuencia.

Texto

El contenido generado por IA puede ser incorrecto.

Tabla

El contenido generado por IA puede ser incorrecto.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.