Seguimiento de activos digitales con Canary Token

Juliana M. Notreni[1], Ninfa M. Zea Cardenas[1], Fabian A. Gibellini[1], German N. Parisi[1], Analia L. Ruhl[1], Leonardo R. Ciceri[1], Marcelo J. Auquer[1], Ileana M. Barrionuevo[1], Federico Bertola[1], Ignacio J. Sanchez Balzaretti[1]

[1] Universidad Tecnológica Nacional – Facultad Regional Córdoba
Maestro M. López esq. Cruz Roja Argentina, Ciudad Universitaria, Córdoba, Argentina
{julinotreni, milyzc, fabiangibellini, germannparisi, analialorenaruhl, leonardorciceri, marcelo.auquer, ilebarrionuevo, fedebertola, ignaciojsb}@gmail.com

Resumen. La fuga de datos a nivel informático existe desde que existen las computadoras en los lugares de trabajo. En los últimos años, considerando los costos económicos y no económicos que este tipo de ataques maliciosos internos acarrean, se han propuesto muchos métodos y técnicas para resolver este problema. Entre las razones claves para implementar mecanismos de prevención de pérdida de datos en una organización están la conformidad con regulaciones establecidas y la protección de la propiedad intelectual. Data Loss Prevention (DLP, Prevención de pérdida de datos) surgió como respuesta a buscar soluciones preventivas a los ataques de atacantes internos que tienen como objetivo la fuga de datos. Es importante implementar Data Loss Prevention, pero, como todo lo relacionado con seguridad y privacidad de datos, no es una bala de plata para las fugas de datos. Por esto, todavía existe la necesidad de poder detectar estos tipos de ataque lo más tempranamente posible para poder minimizar los daños y aplicar los respectivos planes de contingencia. A través del seguimiento de archivos con canary token se pretende detectar un ataque de fuga de datos.

Palabras Claves. ciberseguridad, fuga de datos, Data Loss Prevention, prevención, Canary Token, seguimiento de activos digitales.

Tracking Digital Assets with Canary Token

Abstract. Data leakage at the computer level has been present since computers came up in the workplace. In recent years, considering the economic and non-economic costs that this type of malicious insider attack entails, many methods and techniques have been proposed to solve this problem. Among the key reasons to implement data loss prevention mechanisms in an organization are compliance with established regulations and the protection of intellectual property. Data Loss Prevention (DLP, Data Loss Prevention) arose as a response to seeking preventive solutions to attacks by internal attackers whose objective is data leakage. It is important to implement Data Loss Prevention but, like everything

Received May 2024; Accepted June 2024; Published July 2024
https://doi.org/10.24215/15146774e053

ISSN 1514-6774

related to data security and privacy, it is not a silver bullet for data leaks. For this reason, there is still a need to be able to detect these types of attacks as early as possible in order to minimize the damage and apply the respective contingency plans. Through file tracking with a canary token, it is pretended to detect data leak attacks.

Keywords. cybersecurity, data leak, Data Loss Prevention, prevention, Canary Token, tracking of digital assets.

1 Introducción

¿Por qué es importante incrementar los esfuerzos sobre el control de los datos que se manejan? A continuación se describen algunos ejemplos de cuando estos controles han sido superados, convirtiendo a estos eventos en incidentes de fuga de datos.

Uno de los primeros incidentes y más conocidos de fuga de datos es el incidente de Experian (antes conocido como TRW) en 1984, que expuso información personal y financiera de aproximadamente 90 millones de usuarios. Entre los datos de sus usuarios que administraba TRW se incluía la historia crediticia, datos sobre el empleo, detalles bancarios y de préstamo y lo más importante, números de seguridad social. Lo interesante de este incidente es que previo a la fuga de estos datos, hubo una fuga del manual de operaciones del sistema de TRW, lo que debe haber contribuido a que la fuga de datos no fuera detectada de forma instantánea, sino mucho tiempo después [1].

Otro caso en 2014, quizás menos conocido pero no menos importante, es que se reportó el robo de los datos personales de 20 millones de surcoreanos (40% de la población del país). Los clientes de tres compañías coreanas (KB Kookmin Bank, Lotte Card y Nonghyup Bank) fueron afectados por el robo de datos personales cruciales como números de identificación, direcciones y números de tarjetas de crédito. Las tres firmas fueron multadas y recibieron suspensiones de sus operaciones por las autoridades. Desde el robo de los datos, alrededor de medio millón de clientes habían solicitado que se emitieran nuevas tarjetas de crédito. Los detalles de los clientes fueron desvelados por un trabajador de Korea Credit Bureau, una compañía que ofrece servicios de gestión de riesgos y detección de fraudes. El trabajador, que tenía acceso a varias bases de datos en la empresa, copió durante un año y medio datos en secreto en una unidad externa [2][3].

Como último ejemplo de fuga de datos, en 2016 a un reportero se le mostraron muestras de hasta 24.500 páginas de datos altamente técnicos sobre el submarino Scorpene, un diseño avanzado no nuclear que ha sido exportado por la firma francesa DCNS a varios países. Los documentos incluían dibujos altamente técnicos, especificaciones y descripciones de capacidad operativa de las características de sigilo del submarino; firmas de ruido a diferentes velocidades; alcance, resistencia, profundidades de buceo, datos magnéticos e infrarrojos. Se cree que un subcontratista francés copió los datos de DCNS en Francia en 2011 y la llevó a "un país del sudeste asiático" (según Reuters, India). Después de una disputa con su empleador, el

subcontratista fue despedido, pero los datos se dejaron en una computadora de la empresa. La información se envió a la oficina central de la compañía en Singapur, y en abril de 2013 los datos se colocaron en un servidor. No está claro, dijo The Weekend Australian, cuánto tiempo residieron los datos en el servidor o si algún servicio de inteligencia extranjero obtuvo los datos. Pero el paquete de datos completo se copió en un disco y se envió a una persona en Sydney, Australia, quien, al darse cuenta de la importancia, lo copió en un disco cifrado, destruyó el original y lo almacenó en un archivador cerrado con llave durante más de dos años. Más recientemente, el hombre mostró muestras de los datos a un reportero queriendo demostrar que existía una grave violación de seguridad y que Francia ya había perdido el control de los datos secretos de tales submarinos [4].

Son este tipo de incidentes los disparadores de este proyecto, la fuga de datos existe desde el comienzo de los tiempos de las primeras computadoras.

El objetivo de este paper es introducirnos en la línea de seguridad, específicamente sobre la trazabilidad de datos sensibles por parte de quienes deben tener control y resguardo sobre los mismos.

2 Estado actual sobre seguimiento de documentos con datos sensibles

La fuga de datos ocurre cuando datos sensibles son revelados a partes no autorizadas, ya sea intencionalmente o no. Esto puede representar una amenaza a una organización, ya que la pérdida de datos o confidencialidad puede impactar severamente su reputación y la de sus clientes y empleados; además de que otras organizaciones puedan tomar ventaja de esto.

En algunos casos, el impacto de estas fugas de datos puede superar las fronteras digitales llevando al cierre de dichas organización o inclusive llegar a extremos de generar crisis políticas como fue el caso de WikiLeaks [5].

De acuerdo con un reporte de IBM y el Ponemon Institute basado en 537 casos en 17 países y 17 industrias diferentes, el costo de una fuga de datos en 2021 en promedio fue de 4,24 millones de dólares (un diez por ciento superior respecto del año anterior) [6].

En cuanto a proyecciones para años futuros se puede mencionar:

- Según Cisco, para el 2023 habrá tres veces más dispositivos conectados a la red que humanos [7].
- El mundo almacenará 200 zetabytes (2e14 GB) de datos para el 2025, según Ventures. Estos datos incluyen tanto datos almacenados en infraestructuras públicas como privadas, nubes públicas como privadas, data centers, dispositivos personales y dispositivos IoT [8].

En los últimos años, considerando los costos económicos y no económicos que este tipo de ataques maliciosos internos acarrean, se ha reconocido y visibilizado el desafío de lidiar con ellos y se han propuesto muchos métodos y técnicas para resolver este problema.

Entre las razones claves para implementar mecanismos de prevención de pérdida de datos están la conformidad con regulaciones establecidas y la protección de la propiedad intelectual [9].

En la actualidad, muchas organizaciones y compañías están bajo la supervisión de regulaciones gubernamentales y de la industria que imponen controles sobre la información en general y la información del ámbito privado de las personas en particular. Las regulaciones o normas que una organización debe acatar dependen del ámbito, país o estado donde se desempeñe dicha organización. Algunos ejemplos de normas o regulaciones son:

- HIPAA (Health Insurance Portability and Accountability Act, Ley de Portabilidad y Responsabilidad de Seguros Médicos, en español) [10].
- PCI-DSS (Payment Card Industry Data Security Standard, Estándar de seguridad de datos de la industria de tarjetas de pago, en español), diseñada para que todas las compañías acepten, procesen, almacenen o transmitan datos relacionados a tarjetas de crédito de forma segura [11].
- GDRP (General Data Protection Regulation, European Data Protection Regulation) [12].

Además, muchos estados han aprobado leyes que exigen a las organizaciones que notifiquen a los consumidores cuando su información personal pueda haber sido expuesta [13].

Para muchas compañías, la propiedad intelectual puede ser más valiosa que los activos físicos. Como resultado, para muchas empresas, el establecer políticas y mecanismos que protejan contra la pérdida o robo de propiedad intelectual es crítico para proteger la marca y mantener la competitividad.

Data Loss Prevention (DLP, Prevención de pérdida de datos, en español) surgió como respuesta a buscar soluciones preventivas a los ataques de atacantes internos que tienen como objetivo la fuga de datos [14].

De acuerdo con el NIST (National Institute of Standards and Technology), para prevenir la fuga de información es necesario considerar los siguientes aspectos esenciales [15]:

- Definir políticas de uso de datos, reportes de incidentes de pérdidas de datos y establecimiento de capacidades de respuesta a incidentes para habilitar acciones correctivas y remediar violaciones.
- Definir la sensibilidad de los datos, creación de un inventario de datos sensibles y localización de dónde están siendo almacenados, administración del borrado de datos.
- Monitorear el uso de datos sensibles y entendimiento de patrones de uso de dichos datos.
- Asegurar el cumplimiento de las políticas de seguridad de manera proactiva para prevenir que los datos sensibles salgan de la empresa.

Por su parte, Kostadinov en su artículo Data Loss Protection (DLP) for ICS/SCADA, explica los tres componentes fundamentales de DLP [16]:

- Identificar la información valiosa.
- Mantener seguimiento de las transmisiones de esa información.
- Prevenir acceso no autorizado.

Por último, DLP distingue entre tres estados principales de los datos, requiriendo diferentes técnicas de prevención para cada uno de ellos [15] [17]:

- Data-At-Rest (datos en almacenamiento en computadoras).
- Data-In-Use (cualquier dato con el que el usuario esté interactuando).
- Data-In-Motion (datos siendo enviados a través de una red).

Entre las tecnologías usadas para dar protección a los datos en sus diferentes estados, se pueden encontrar entre otras: Intrusion Detection Systems (IDS) [18], Intrusion Prevention Systems (IPS), antimalwares, firewalls, actualizaciones de software y Security Information Event Management (SIEM) [19] [20].

La mayoría de las soluciones de DLP fueron construidas hace diez o quince años sin poder tener en cuenta las particularidades del mundo actual respecto al teletrabajo, BYOD (Bring Your Own Device, tendencia a que los empleados usen sus propios dispositivos personales con fines laborales) y la nube. A pesar de que se han desarrollado nuevas características y mejoras durante todo este tiempo, existen muchos puntos ciegos con los que las actuales soluciones de DLP no pueden lidiar. A continuación, se listan algunas de estas limitaciones que pueden ser usadas para eludir DLP [21]:

- Encriptado del archivo antes de enviarlo. Si se encripta el archivo, DLP no podrá leerlo. Aquí se tiene que tomar una decisión si se bloquea o no la transmisión de este tipo de archivos.
- Tomar fotografías y enviarlas. Mientras que algunas soluciones DLP soportan capacidad OCR (Optical Character Recognition, Reconocimiento óptico de caracteres, en español) en tiempo real, la limitación es que solo se soporta al nivel del gateway y no al nivel del terminal. Si alguien toma capturas de pantalla y las embebe en un archivo de ofimática, por ejemplo, casi con seguridad eludirá la solución DLP.
- Copiar datos hacia teléfonos móviles (Android) usando un cable. Las soluciones DLP han sido tradicionalmente débiles en controlar la transferencia de datos a teléfonos Android conectados vía el puerto USB.
- Usar Linux o sistemas Mac o virtualización. La mayoría de las soluciones DLP no soportan sistemas Mac o Linux. Incluso usando Windows, un atacante puede instalar una plataforma de virtualización para crear una máquina virtual con Linux y enviar datos de manera exitosa, dado que la solución DLP a nivel de terminal no podrá monitorear las actividades dentro de la máquina virtual.
- Usar el modo incógnito del navegador o el modo seguro de Windows. El modo incógnito del navegador web es un punto ciego de la mayoría de las soluciones DLP e iniciar Windows en modo seguro es otro punto ciego ya que los servicios DLP no trabajan en dicho modo [22].
- Insertar datos en archivos grandes (más de 20 MB). A medida que el tamaño de archivo crece, se sobrecargan los recursos del terminal que DLP requiere, por lo que la mayoría de las soluciones DLP no monitorean archivos de tamaño considerable.

- Capturar la pantalla usando dispositivos móviles o cámara. DLP no puede detectar lo que pasa fuera del sistema.

Dado que ningún sistema es 100% seguro y por la existencia de limitaciones en Data Loss Prevention, es absolutamente necesario que las organizaciones estén preparadas para gestionar las posibles fugas de datos que eventualmente se produzcan. Para poder gestionarlas es necesario primero identificarlas, lo cual conlleva tener trazabilidad de los datos sensibles (y de los archivos que los contengan).

Los puntos expuestos anteriormente, como los ejemplos de fuga de datos, la posible solución con Data Loss Prevention y sus limitantes son la punta del iceberg al problema de fuga de datos y por lo cual surge este proyecto, el cual propone explorar la implementación de un mecanismo, específicamente un framework, que le permita a una organización tener una visibilidad más inmediata sobre algunos eventos de fugas de información con el consiguiente incremento en su capacidad de reacción para ejecutar planes de contingencia previamente definidos vinculados a los riesgos de fuga de dicha información.

Se explorará el concepto de canary tokens para lograr la trazabilidad de archivos que contienen datos sensibles. Los canary tokens en seguridad informática a menudo aluden al concepto del canario en una mina de carbón donde los pájaros eran una señal de advertencia temprana de que el peligro estaba cerca. Si los canarios de la mina morían, servía como indicación de que los mineros debían salir de inmediato porque los canarios eran más sensibles a los gases peligrosos que los humanos. Actualmente esta idea se traslada al mundo digital, utilizando estos “canarios digitales” para ser alertado en el caso de que surja alguna actividad no deseada.

Reale et al definen que los softwares que implementan canary tokens distribuyen tokens; un token es un identificador único generado de forma aleatoria y puede ser ubicado en URLs o en otras propiedades como hostnames. Cuando la URL o dicha propiedad es requerida, se alerta al propietario del token proveyéndolos de información acerca de este evento [23].

Si bien de acuerdo con la definición anterior, el canary token puede ser ubicado en URLs o hostnames, puede darse el caso, de acuerdo a la especificación del formato del archivo donde el mismo será embebido, que no sea factible su transmisión mediante el protocolo HTTP [24] y por ende ser necesario su inserción en otro tipo de estructura y la utilización de un protocolo diferente para su transmisión.

Actualmente, una de las plataformas más conocidas para generar canary tokens y de distintos tipos es canarytokens.org creada por la organización Thinkst y de código abierto [25] [26]. Esta plataforma cuenta diversos tipos de tokens, entre ellos se puede mencionar Token DNS, Claves de AWS (notifica cuando alguien usa esas credenciales), Token log4shell [27] (si alguna librería es vulnerable a la vulnerabilidad de log4shell), etc. Estas plataformas, para el caso de documentos, lo que generan es el documento con el token ya inyectado y en algunos casos el archivo se puede seguir completando y se envía una notificación a una dirección de correo electrónico o un webhook cuando el documento es abierto.

También se utilizan estas plataformas a los fines de detectar vulnerabilidades de configuración inadecuada de servidores, en donde se puede obtener información interna como IPs privadas, información de infraestructura en la nube y variables de entorno, entre otros datos. Entre las plataformas más utilizadas se encuentran: dnslog.cn, webhook.site, interact.projectdiscovery.io, pingb.in, swin.es, ceye.io, requestbin.net, beeceptor.com, y la herramienta Burpcollaborator que se encuentra disponible en la herramienta Burpsuite [28]. Una vulnerabilidad que suele utilizar canary tokens en su fase de descubrimiento es aquella llamada Server Side Request Forgery (SSRF) [29].

Como se expuso anteriormente Data Loss Prevention no asegura que no existan fugas de datos es necesario contar, además de herramientas DLP también con herramientas que permitan detectar este tipo de ataques lo más tempranamente posible.

Actualmente, la plataforma de canary token permite trabajar de a un documento pero ¿Qué pasa cuando se necesita tener rastreabilidad de cientos o miles de archivos? como es el caso de las organizaciones, es por esto que esta línea de investigación, incluida en seguridad informática, pretende ampliar el uso de canary tokens y que también estos puedan ser considerados desde la concepción de cualquier proyecto de software, por ejemplo, ¿Es necesario tener trazabilidad de todos los documentos generados por una organización? ¿Cómo identificamos los que necesitan ser rastreados o monitoreados de los que no? ¿Qué documentos tienen que ser rastreados? ¿Qué datos es necesario recopilar de cada documento ya rastreado? Si estas interrogantes son contestadas afirmativamente entonces estamos ante casos en los que sería interesante considerar implementar canary tokens en varios documentos. Es por esto que uno de los puntos de este proyecto es considerar tener rastreabilidad sobre documentación masiva.

El objetivo de este proyecto es minimizar los daños ante una fuga de datos, a través, del seguimiento de datos (archivos) alertando cuando estos sean abiertos desde orígenes desconocidos y no autorizados de forma que la organización pueda implementar sus respectivos planes de contingencia antes estos eventos.

3 Grado de avance

Actualmente el proyecto se encuentra en etapa de investigación exploratoria, lo que sí se ha identificado la necesidad de los siguientes puntos para alcanzar los objetivos propuestos:

- Desarrollar un mecanismo que permita inyectar en diferentes tipos de archivos (pdf, docx, xlsx, etc.) un canary token que facilite la obtención de información acerca de las circunstancias en las que el archivo es consultado, de manera de tener visibilidad respecto de si se ha consumado una fuga de información.
- Recolectar la información recibida de los documentos generadores a partir de esta biblioteca y emitir las alertas correspondientes.

De esta forma se puede comenzar a responder las interrogantes planteadas previamente:

- ¿Es necesario tener trazabilidad de todos los documentos generados por una organización? ¿Cómo identificamos los que necesitan ser rastreados o monitoreados de los que no?
- ¿Qué documentos tienen que ser rastreados?
- ¿Qué datos es necesario recopilar de cada documento ya rastreado?

Para poder identificar el mecanismo que permita la inyección de canary tokens es necesario investigar/estudiar el estándar ECMA-376. ECMA-376 especifica una familia de esquemas XML, denominados colectivamente Office Open XML, que definen el formato XML. vocabularios para procesamiento de textos, hojas de cálculo y documentos de oficina de presentación, así como el empaquetado de documentos ofimáticos que se ajusten a estos esquemas. El objetivo es permitir la implementación de los formatos Office Open XML mediante el más amplio conjunto de herramientas y plataformas, fomentando la interoperabilidad entre aplicaciones de productividad de oficina y sistemas de línea de negocio, así como apoyar y fortalecer el archivo y conservación de documentos, todo ello de forma totalmente compatible con los existentes documentos de Microsoft® Office [30].

En base a lo analizado se han identificado tres componentes que van a permitir responder los interrogantes planteados:

- Un inyector de tokens a documentos (Figura 1.).
- Gestor del inyector de documentos (Figura 2.).
- Visualizador de datos recibidos que envían los documentos tokenizados (Figura 3).

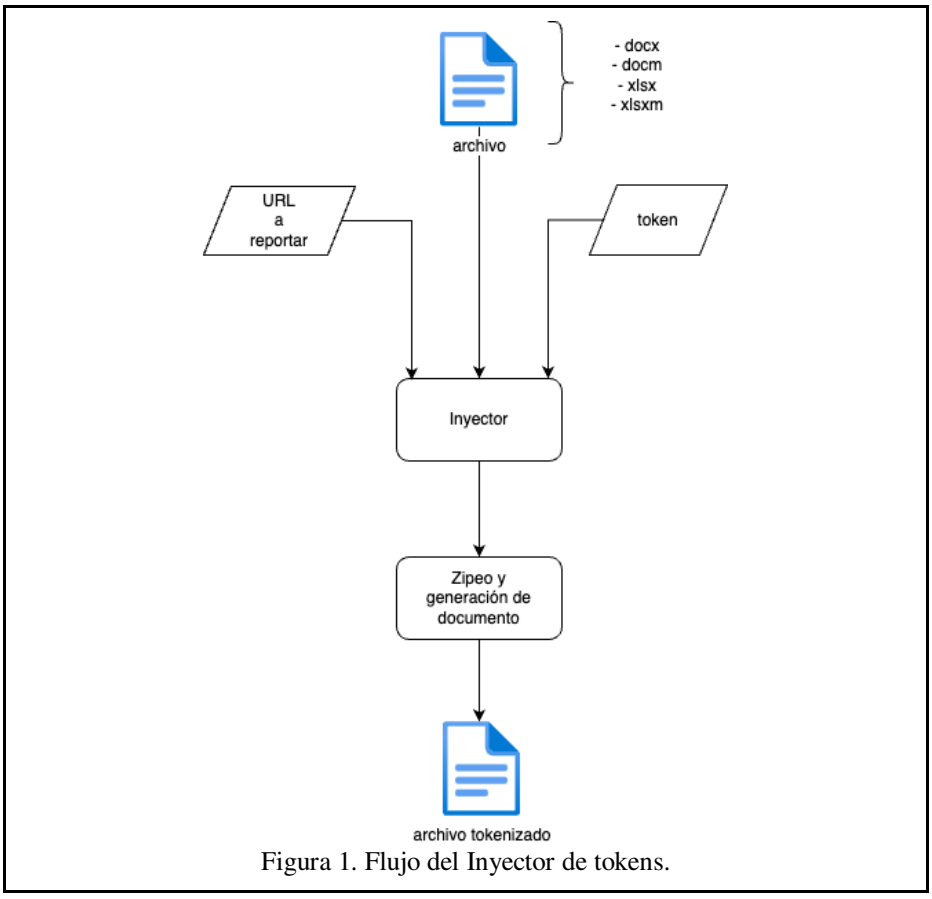

Figura 1. Flujo del Inyector de tokens.

El inyector será el encargado de insertar el token y la url a que se tiene que reportar en los documentos. Como existen diferentes tipos de archivos es necesario acotar el alcance de los mismos ya que cada tipo de documento requiere cierta investigación previa para poder insertar un canary token.

Inicialmente se ha decidido centrarse en cuatro tipos de archivos:

- docx (Documento Word)
- doxm (Documento Word habilitado para macro)
- xlsx (Documento Excel)
- xlxm (Documento Excelhabilitado para macro)

Para llevar a cabo este inyector se ha decidido utilizar python como lenguaje de programación, debido a su portabilidad. Por otro lado, se viene identificado en los tipos de documentos mencionados donde se podría anexar el canary token teniendo en cuenta que están siendo trabajados como XML, basándose en el estándar mencionado anteriormente.

La siguiente estructura (Tabla 1.) es la estructura de un de un documento .docx o un .docm. Para confirmar la factibilidad de la inyección de una URL de reporte de datos como también para identificar qué datos se pueden obtener de dichos documentos se planteó una prueba de concepto para verificar o no el envío de datos a una URL determinada.

```
LEER_docx
├── [Content_Types].xml
├── _rels
├── docProps
│   ├── app.xml
│   └── core.xml
└── word
    ├── _rels
    │   ├── document.xml.rels
    │   └── footer2.xml.rels
    ├── document.xml
    ├── endnotes.xml
    ├── fontTable.xml
    ├── footer1.xml
    ├── footer2.xml
    ├── footer3.xml
    ├── footnotes.xml
    ├── header1.xml
    ├── header2.xml
    ├── header3.xml
    ├── settings.xml
    ├── styles.xml
    ├── theme
    │   └── theme1.xml
    └── webSettings.xml
```

Tabla 1. Estructura de un documento de texto .docx o .docm, especificado por ECMA-376

A partir de la Tabla 3, se pueden describir los documentos más relevantes

- [Content_Types].xml: Elemento de tipo de contenido.
- /_rels/.rels : Elemento de relación de paquete.
- /docProps/app.xml: Parte de propiedades definidas por la aplicación.
- /docProps/core.xml: Parte de propiedades core.
- /word/document.xml: Parte del documento principal
- /word/footnotes.xml : Footnotes part

A partir de estos se pudo identificar los archivos necesarios intervinientes en el envío de datos a un servidor arbitrario. Office Open XML impone restricciones a las

relaciones. Las relaciones en Office Open XML son explícitas o implícitas. Para una relación explícita, se hace referencia a un recurso desde el XML de una parte fuente utilizando el atributo Id de una etiqueta Relationship. Algunas relaciones deben ser explícitas. Todas las demás relaciones son implícitas. La sintaxis para especificar una relación implícita varía entre los tipos de etiquetas.

En la prueba de concepto, se busca recibir los siguientes datos:

- URL: Dirección web a la cual reportará los eventos que se ejecuten sobre el archivo.
- SOFTWARE: tipo de documento, .docx en este caso.
- CAMPAIGN_ID: identificador de prueba de concepto.
- TOKEN: token asociado al documento.
- ORIGINAL_FILENAME: nombre del archivo en cuestión.
- SCHEMANAME: un valor definido “ignore”.
- INTERACTION: Evento que se ejecuta sobre el documento por el cual se está recibiendo información. Para iniciar el valor de este siempre será “opened_file” que hace referencia a que el documento ha sido abierto.

Para poder lograr la recepción de datos cuando un usuario abre un documento se logró identificar el archivo sobre el que se tiene que trabajar (Tabla 2.). De esta forma, cuando un usuario cualquiera abra el documento en cuestión esto desencadenará una llamada HTTP GET al servidor definido por el campo URL. Esta modificación se lleva a cabo en el archivo /word/_rels/footer2.xml.rels.

```
<Relationships
xmlns="http://schemas.openxmlformats.org/package/2006/relationships">
   <Relationship Id="rId1"

Type="http://schemas.openxmlformats.org/officeDocument/2006/relationships/i
mage"

Target="URL?d=s__SOFTWARE__c__CAMPAIGN_ID__t__TOKEN__o__O
RIGINAL_FILENAME__s__SCHEMANAME__i__INTERACTION"
      TargetMode="External" />
</Relationships>
```

Tabla 2. Etiqueta XML anexada en un documento .docx para poder insertar los tokens.

Considerando que también se incluyó en esta PoC a los documentos habilitados para macros (.docm), para estos fue necesario cambiar el documento y alguno de los datos que se esperan recibir (Tabla 3.). Esta modificación se lleva a cabo en el archivo /word/document.xml.

```
<w:r w:rsidRPr="001973E5">
   <w:rPr>
    <w:color w:val="FFFFFF" w:themeColor="background1" />
    <w:sz w:val="12" />
    <w:szCs w:val="12" />
   </w:rPr>
   <w:t>

replace_campaign_id|replace_token|replace_schema_name|replace_filename|repl
ace_url|
   </w:t>
 </w:r>
```

Tabla 3. Etiqueta XML anexada en un documento .docm para poder insertar los tokens. en el archivo ./word/_rels/footer2.xml.rels

Para el caso de los documentos excel, la estructura de los archivos especificada por la ECMA-376 cambia (Tabla 4.)

```
LEER_xlsx
├── [Content_Types].xml
├── _rels
├── docProps
│   ├── app.xml
│   └── core.xml
└── xl
    ├── _rels
    │   └── workbook.xml.rels
    ├── drawings
    │   ├── _rels
    │   │   └── drawing1.xml.rels
    │   └── drawing1.xml
    ├── printerSettings
    │   └── printerSettings1.bin
    ├── sharedStrings.xml
    ├── styles.xml
    ├── theme
    │   └── theme1.xml
    ├── workbook.xml
    └── worksheets
        ├── _rels
```

```
└── sheet1.xml.rels
└── sheet1.xml
```

Tabla 3. Etiqueta XML anexada en un documento .docm para poder insertar los tokens. en el archivo ./word/_rels/footer2.xml.rels

A diferencia de los tipos de archivos anteriores (.docx y docm), en un .xlxs podemos identificar:

- /xl/workbook.xml: Parte del workbook
- /xl/_rels/workbook.xml.rels: Elemento de relaciones
- /xl/sharedStrings.xml: Parte de las tablas de caracteres compartidas.
- /xl/worksheets/sheet1.xml…./xl/worksheets/sheetN.xml: parte de hojas de trabajo.

Para el caso de los documentos de tipo .xlsx se considera el mismo contenido de la tabla 2 pero dentro de otro archivo llamado /xl/drawings/_rels/drawing1.xml.rels. de la misma forma que para los documentos habilitados para macros .xlsm existe otro archivo interviniente denominado /xl/sharedStrings.xml.

Todo lo mencionado anteriormente constituye el core o núcleo del inyector de tokens a documentos. El objetivo es centralizar la inyección de canary token usado, usando dicho inyector que será llamado cada vez que se necesite insertar un token en algún documento.

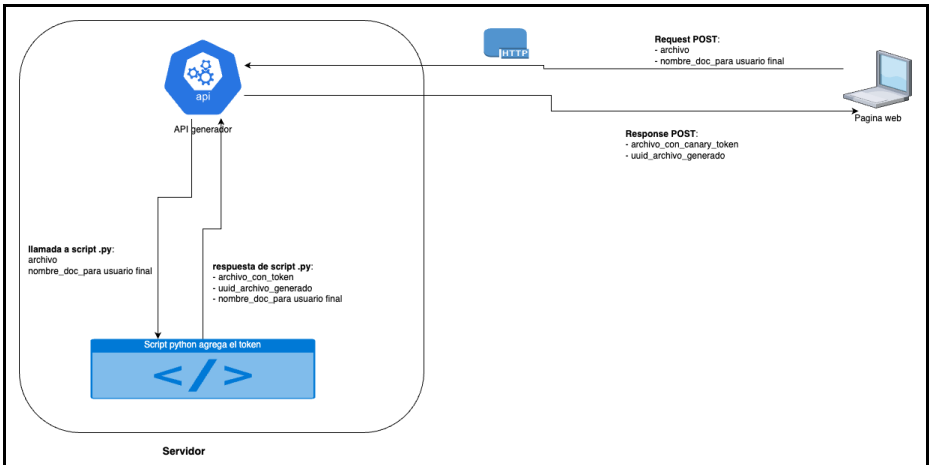

Figura 2. Gestor de documentos e Inyector de tokens

Todo esto estará disponible a través de un sitio web, de similar forma en la que actualmente está disponible la herramienta Canary tokens (https://canarytokens.org/generate).

Por otro lado, tenemos el lector de datos (Figura 3) que nos permitirá visualizar los datos recolectados a partir de los documentos que tiene un token inyectado como también procesar estos datos y obtener estadísticas.

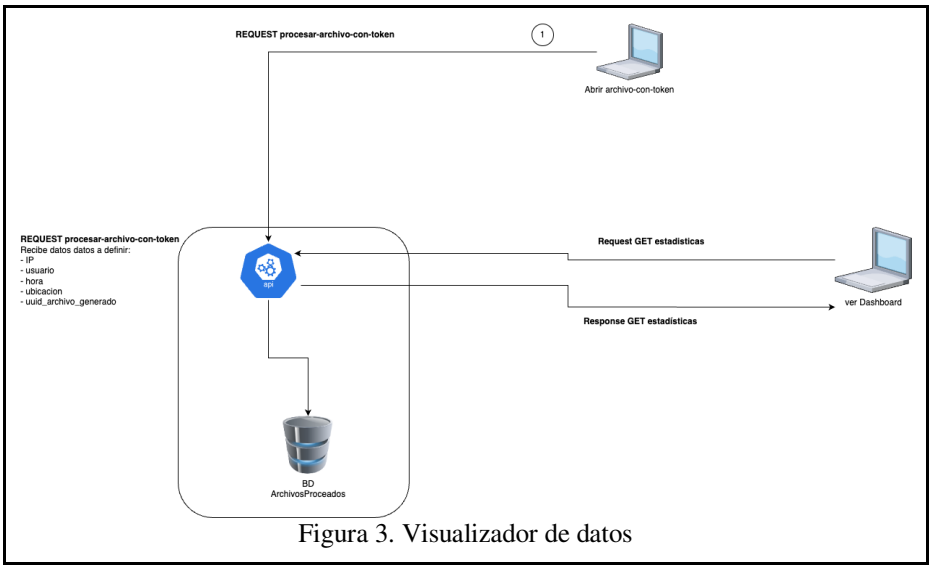

Figura 3. Visualizador de datos

4 Conclusión

Si bien la fuga de datos digitales es un problema que se acarrea desde siempre, los mecanismos para enfrentarlos enfocados con vehemencia en este mal son recientes como Data Loss Prevention y los Canary token.

El fin de este trabajo es lograr un mecanismo que entre sus cualidades está la portabilidad, de esta forma se podría aplicar tanto a documentos ya existentes como a documentos generados en cualquier sistema. Además de ser independiente del sistema operativo sobre el que se trabaja día a día y sobre el que se ejecuta el sistema que genera los documentos en cuestión.

Si bien los avances hasta ahora van siendo teóricos y actualmente se está desarrollando la prueba de concepto, se ha tenido en cuenta factores como que la portabilidad con un horizonte a que el código generado pueda llegar a ser eventualmente código abierto, de forma tal que permita generar mayor conocimiento sobre estos mecanismos de protección de datos en archivos tan usados como excel y word o cualquier otro documento que implemente el estándar ECMA-376

References

1. Ozkaya, E. Julio 2021. The history of data breaches. https://www.erdalozkaya.com/the-history-of-data-breaches/#1984_-_The_TRW_data_breach, última visita 07/04/2022.
2. Yan, Sophia & Kwon, K. J.. (2014). Massive data theft hits 40% of South Koreans. https://money.cnn.com/2014/01/21/technology/korea-data-hack/, última consulta: 21/5/2022.
3. AFP. (Febrero 2014). South Korean Credit Card Firms Punished for Data Leak. https://www.securityweek.com/south-korean-credit-card-firms-punished-data-leak, última visita: 21/5/2022.
4. (2016). Submarine Data Leak Roils Three Governments. https://www.defensenews.com/naval/2016/08/26/submarine-data-leak-roils-three-governments/, última visita el 21/5/2022.
5. Tahboub, Radwan & Saleh, Yousef. (2014). Data Leakage/Loss Prevention Systems (DLP). International Journal of Information Systems. 1. 13-19. 10.1109/WCCAIS.2014.6916624.
6. Tunggal, A. (Mayo 2022) What is the Cost of a Data Breach in 2022?. [https://www.upguard.com/blog/cost-of-data-breach].
7. Cisco. (Febrero 2020) Cisco Annual Internet Report Forecasts 5G to Support More Than 10% of Global Mobile Connections by 2023.
8. The 2020 Data Attack Surface Report. Arcserve Tape Backup Whitepaper. Última visita: https://1c7fab3im83f5gqiow2qqs2k-wpengine.netdna-ssl.com/wp-content/uploads/2020/12/ArcserveDataReport2020.pdf.
9. Forcepoint. Forcepoint Data Loss Prevention (DLP). Protección de datos en un mundo sin perímetros. https://www.forcepoint.com/sites/default/files/resources/brochures/brochure-dlp-es.pdf, última visita: 18/4/2022.
10. The HIPAA Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html, última visita 15/04/2022.
11. PCI Security Standards. (Marzo 2022). Payment Card Industry. Estándar de Seguridad de Datos. Requisitos y Procedimientos de Evaluación. https://www.pcisecuritystandards.org/documents/PCI-DSS-v4_0-LA.pdf?agreement=true&time=1653751557059, última visita: 28/5/2022.
12. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679, última visita 15/04/2022.
13. Data Loss Prevention https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=904672, última visita: 18/4/2022.
14. Papadimitriou, Panagiotis & Garcia-Molina, Hector. (2011). Data Leakage Detection. Knowledge and Data Engineering, IEEE Transactions on. 23. 51 - 63. 10.1109/TKDE.2010.100.
15. National Institute of Standards and Technology NIST. Data Loss Prevention https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=904672, última visita: 19/4/2022.
16. Kostadinov, D. (2020). Data Loss Protection (DLP) for ICS/SCADA. https://resources.infosecinstitute.com/topic/data-loss-protection-dlp-for-ics-scada/, última visita 21/05/2022.
17. The SANS Institute. Securosis, L.L.C. Understanding and Selecting a Data Loss Prevention Solution. https://securosis.com/assets/library/publications/DLP-Whitepaper.pdf, última visita 19/4/2022.
18. SANS. (2017). SANS Institute: Reading Room - Intrusion Detection. https://www.sans.org/readingroom/whitepapers/detection/paper/38165.
19. What is SIEM?. https://www.ibm.com/topics/siem, última visita 15/04/2022].
20. Tahboub, Radwan & Saleh, Yousef. (2014). Data Leakage/Loss Prevention Systems (DLP). International Journal of Information Systems. 1. 13-19. 10.1109/WCCAIS.2014.6916624.

21. Kumar AS. (2021). Data Loss Prevention: DLP limitations and how to bypass?. https://securityfocal.com/data-loss-prevention-dlp-limitations-and-how-to-bypass/, última visita 19/4/2022.
22. (2019). Top 4 Reasons Why You Should Include Behavioral Analysis in DLP | Digital Guardian, Digital Guardian. https://digitalguardian.com/resources/webinar/top-4-reasons-why-youshould-include-behavioral-analysis-dlp
23. Reale A., Zinc B. (2019). Loft: Canarytokens: An old concept for a new world. Scientific and Practical Cyber Security Journal (SPCSJ) 3(1): 66- 68 ISSN 2587-4667 Scientific Cyber Security Association (SCSA)
24. Hypertext Transfer Protocol -- HTTP/1.1. RFC-2616. https://datatracker.ietf.org/doc/html/rfc2616.
25. Canary tokens. Página oficial. https://www.canarytokens.org/generate.
26. Código de Canary tokens. Github. Página oficial. https://github.com/thinkst/canarytokens.
27. R Hiesgen, M Nawrocki, TC Schmidt, M Wählisch. (2022). The Race to the Vulnerable: Measuring the Log4j Shell Incident. arXiv preprint arXiv:2205.02544.
28. Página oficial burpsuite. https://portswigger.net/burp
29. Hacktricks. SSRF (Server Side Request Forgery). https://book.hacktricks.xyz/pentesting-web/ssrf-server-side-request-forgery.
30. ECMA-376 - Ecma International. Office Open XML file format. 5th Edition December 2021. https://ecma-international.org/publications-and-standards/standards/ecma-376/