Framework de detección de fuga de datos por medio de canary tokens

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]

[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}@gmail.com

Resumen

La fuga de datos ocurre cuando información sensible es revelada a partes no autorizadas, ya sea intencionalmente o no. Esto puede representar una amenaza en las organizaciones, debido a 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 sobre 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. Data Loss Prevention (DLP, Prevención de pérdida de datos) surgió como respuesta a buscar soluciones preventivas a los ataques de agresores internos que tienen como objetivo la fuga de datos. A continuación, se presenta un conjunto de herramientas (framework) que permiten tener rastreabilidad de los archivos idóneos de una organización, brindando trazabilidad en todo momento debido a su confidencialidad y sensibilidad de datos. Además se introduce una herramienta de línea de comando, multiplataforma, que permite inyectar canary tokens a archivos desacoplando el uso del framework presentado.

Introducción

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 o inclusive llegar a extremos de generar crisis políticas como fue el caso de WikiLeaks [1].

De acuerdo a 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) [2].

Según Ventures, el mundo almacenará 200 zetabytes (2e14 GB) de datos para el 2025. Esto incluye tanto datos almacenados en infraestructuras y nubes tanto públicas como privadas, Data Centers, dispositivos personales y dispositivos IoT [3].

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 [4].

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) [5].
- 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 [6].
- GDRP (General Data Protection Regulation, European Data Protection Regulation) [7].

Además, muchos Estados han aprobado leyes que exigen a las organizaciones que notifiquen a los consumidores cuando su información personal pueda haber sido expuesta [8].

Para muchas compañías, la propiedad intelectual puede ser más valiosa que los activos físicos. Como resultado, para las empresas, el establecer políticas y mecanismos que

protejan contra la pérdida o robo de propiedad intelectual es crítico para resguardar 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 agresores internos que tienen como objetivo la fuga de datos [9].

De acuerdo con el NIST (National Institute of Standards and Technology), para prevenir la fuga de información es necesario considerar los siguientes aspectos esenciales [10]:

- 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 [11]:

- 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 [10] [12]:

- 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) [13], Intrusion Prevention Systems (IPS), antimalwares, firewalls, actualizaciones de software y Security Information Event Management (SIEM) [14] [15].

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).

El presente trabajo pretende proponer 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.

Para esto se usa 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 software 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 [16].

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 [17] [18]. 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 [19] (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 [20]. Una vulnerabilidad que suele utilizar Canary Tokens en su fase de descubrimiento es aquella llamada Server Side Request Forgery (SSRF) [21].

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

Actualmente, la plataforma de Canary Token permite trabajar de a un documento por vez, pero ¿Qué pasa cuando se necesita tener rastreabilidad de cientos o miles de archivos como es el caso de las organizaciones? 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 éstos 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 a la vez. Es por esto, que uno de los puntos de este proyecto es considerar tener rastreabilidad sobre documentación masiva.

El objetivo de este framework 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.

Metodología

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 del 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.

Avances y Resultados

A partir de una primera etapa se identificaron las siguientes metas:

- 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 se estudió el estándar ECMA-376. Este estándar 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 así como apoyar y fortalecer el archivo y conservación de documentos, todo ello de forma totalmente compatible con los existentes documentos de Microsoft® Office [22].

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).

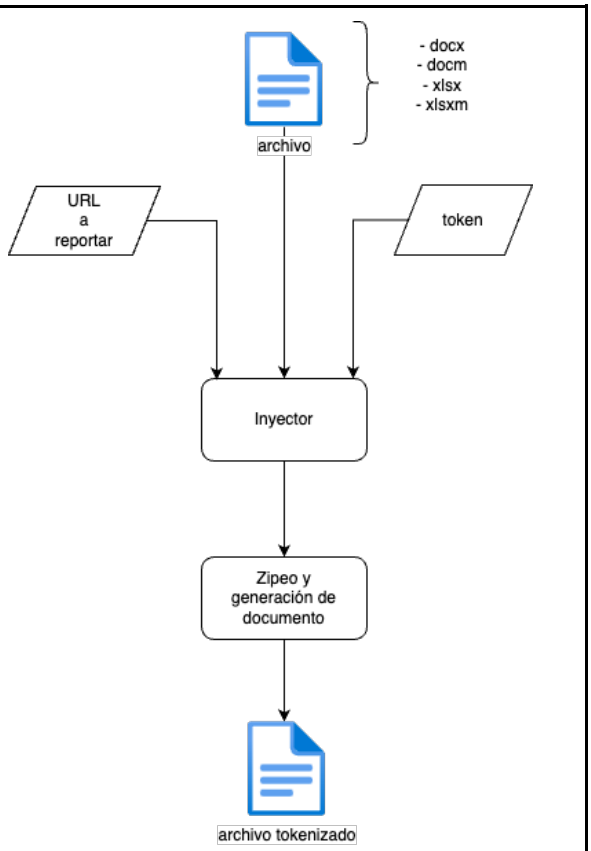

Figura 1. Flujo del Inyector de tokens.

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)
- docm (Documento Word habilitado para macro)
- xlsx (Documento Excel)
- xlsm (Documento Excel habilitado para macro)
- pdf
- exe (ejecutables windows)

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.

El objetivo es centralizar la inyección de Canary Token usado, usando dicho un inyector que será llamado cada vez que se necesite insertar un token en algún documento.

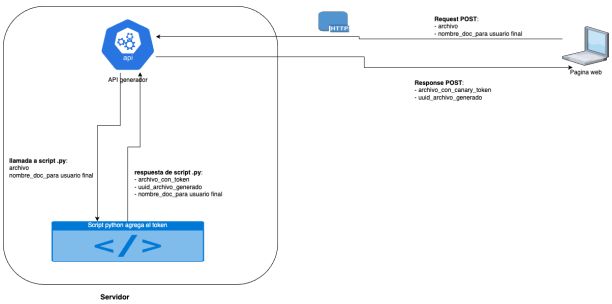

Figura 2. Gestor de documentos e Inyector de tokens

Esta funcionalidad se pondrá a disposición a través de un sitio web.

Por otro lado, se tiene el cuenta con un Visualizador de datos (Figura 3) que nos le permitirá a las organizaciones visualizar los datos la información recolectada a partir la apertura de los documentos que tiene un token inyectado, como también procesar estos datos y obtener estadísticas.

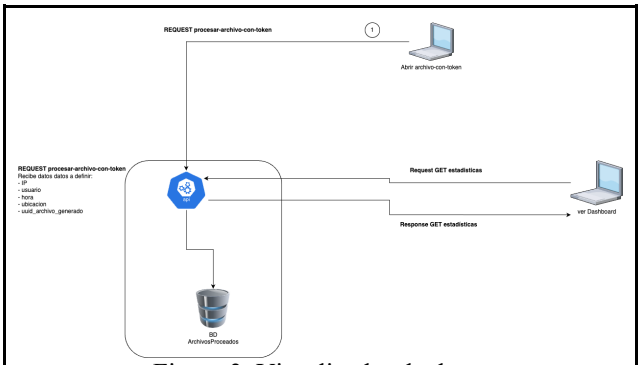

Figura 3. Visualizador de datos

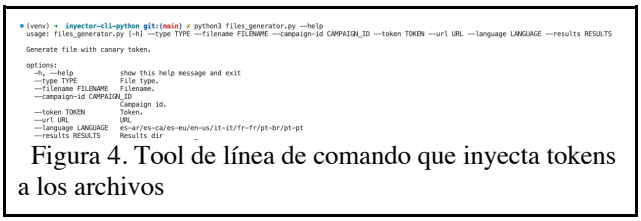

Figura 4. Tool de línea de comando que inyecta tokens a los archivos

El componente principal que inyecta los token a los archivos es una tool de línea de comando, esta tool representa el core de este proyecto (Figura 4).

Conclusiones

Si bien la fuga de datos digitales es un problema que se acarrea desde los orígenes, 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.

Para el desarrollo de la tool que inyecta tokens se han 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. Por esto se ha desarrollado un script en Python, el cual permite su ejecución tanto en Linux, Windows o MacOS.

La herramienta de línea de comando presenta su propio menú de ayuda de forma que también se pueda ejecutar de forma independiente.

Como trabajo futuro de investigación queda seguir explorando soporte a nuevos tipos de archivos que actualmente soporte doc, docm, xlsx, xlsm, pdf, exe (ejecutables windows). Esto se podría extender a pptx y/o pptm, formatos actualmente usados para presentaciones.

Referencias

[1] Tahboub, Radwan & Saleh, Yousef. (2014). Data Leakage/Loss Prevention Systems (DLP). International Journal of Information Systems. 1. 13-19. 10.1109/WCCAIS.2014.6916624.

[2] Tunggal, A. (Mayo 2022) What is the Cost of a Data Breach in 2022?. [https://www.upguard.com/blog/cost-of-data-breach].

[3] 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.

[4] 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.

[5] The HIPAA Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html, última visita 15/04/2022.

[6] 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.

[7] Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679, última visita 15/04/2022.

[8] Data Loss Prevention https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=904672, última visita: 18/4/2022.

[9] Papadimitriou, Panagiotis & Garcia-Molina, Hector. (2011). Data Leakage Detection. Knowledge and Data Engineering, IEEE Transactions on. 23. 51 - 63. 10.1109/TKDE.2010.100.

[10] 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.

[11] 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.

[12] 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.

[13] SANS. (2017). SANS Institute: Reading Room - Intrusion Detection. https://www.sans.org/readingroom/whitepapers/detection/paper/38165.

[14] What is SIEM?. https://www.ibm.com/topics/siem, última visita 15/04/2022].

[15] Tahboub, Radwan & Saleh, Yousef. (2014). Data Leakage/Loss Prevention Systems (DLP). International Journal of Information Systems. 1. 13-19. 10.1109/WCCAIS.2014.6916624.

[16] 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)

[17] Canary tokens. Página oficial. https://www.canarytokens.org/generate.

[18] Código de Canary tokens. Github. Página oficial. https://github.com/thinkst/canarytokens.

[19] R Hiesgen, M Nawrocki, TC Schmidt, M Wählisch. (2022). The Race to the Vulnerable: Measuring the Log4j Shell Incident. arXiv preprint arXiv:2205.02544.

[20] Página oficial burpsuite. https://portswigger.net/burp

[21] Hacktricks. SSRF (Server Side Request Forgery). https://book.hacktricks.xyz/pentesting-web/ssrf-server-side-request-forgery.

[22] ECMA-376 - Ecma International. Office Open XML file format. 5th Edition December 2021. https://ecma-international.org/publications-and-standards/standards/ecma-376/