Minimizando daños de fuga de datos por medio de archivos tokenizados y su trazabilidad de aperturas

Minimizing data leak damage through tokenized files and their opening traceability

Presentación: 23/08/2024

Juliana NOTRENI

Universidad Tecnológica Nacional – Facultad Regional Córdoba
julinotreni@gmail.com

Milagros ZEA CARDENAS
Universidad Tecnológica Nacional – Facultad Regional Córdoba
milyzc@gmail.com

Germán PARISI
Universidad Tecnológica Nacional – Facultad Regional Córdoba
germannparisi@gmail.com

Fabián GIBELLINI
Universidad Tecnológica Nacional – Facultad Regional Córdoba
fabiangibellini@gmail.com

Analía RUHL
Universidad Tecnológica Nacional – Facultad Regional Córdoba
analialorenaruhl@gmail.com

Marcelo AUQUER
Universidad Tecnológica Nacional – Facultad Regional Córdoba
marcelo.auquer@gmail.com

Ileana BARRIONUEVO
Universidad Tecnológica Nacional – Facultad Regional Córdoba
ilebarrionuevo@gmail.com

Federico BERTOLA
Universidad Tecnológica Nacional – Facultad Regional Córdoba
fedebertola@gmail.com

Leonardo CICERI
Universidad Tecnológica Nacional – Facultad Regional Córdoba
leonardociceri@gmail.com

Ignacio SÁNCHEZ BALZARETTI
Universidad Tecnológica Nacional – Facultad Regional Córdoba
ignaciojsb@gmail.com

Resumen

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 pueden 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 atacantes internos que tienen como objetivo la fuga de datos.

En el siguiente trabajo se presenta una alternativa para tener rastreabilidad de los archivos idóneos de una organización y los cuales requieren tener una trazabilidad 24x7 debido a su confidencialidad y sensibilidad de datos. La solución pensada está basada en el concepto de Canary Token.

Palabras clave: Seguridad, Fuga de Datos, Prevención de pérdida de Datos (DLP), Canary Token.

Abstract

Data breaches occur when sensitive information is disclosed to unauthorized parties, whether intentionally or not. This can pose a significant threat to an organization, as the loss of data or confidentiality can severely impact its reputation, as well as that of its clients and employees. Additionally, other organizations may take advantage of such breaches.

In some cases, the impact of these data breaches can extend beyond digital borders, leading to the closure of the affected organization or even escalating to the point of creating political crises, as was the case with WikiLeaks. Data Loss Prevention (DLP) emerged as a response to seek preventive solutions against internal attackers whose objective is data leakage.

In the following work, an alternative is presented to ensure traceability of the sensitive files within an organization, which require 24/7 monitoring due to their confidentiality and data sensitivity. The proposed solution is based on the concept of Canary Token.

Keywords: Security, Data Leak, Data Loss Prevention (DLP), Canary Token.

Introducción

Un incidente de fuga de información sensible se produce cuando dicha información es accesible para personas no autorizadas, ya sea de forma deliberada o accidental. Este tipo de incidentes puede poner en riesgo a una organización. ya que la pérdida de datos o la falta de confidencialidad puede perjudicar gravemente su reputación, así como la de sus clientes y empleados (Yan y Kwon, 2014) (AFP, 2014) (sTAFF, 2016). Además, este tipo de situaciones puede ser explotado por otras organizaciones.

En ciertos casos, las consecuencias de estas fugas de información van más allá del entorno digital, pudiendo ocasionar el cierre de la organización afectada o incluso desencadenar crisis políticas como lo demostró el caso de WikiLeaks (Tahboub et al., 2014:13-19).

Un estudio realizado por IBM y el Pokemon Institute, que analizó 537 casos en 17 países y 17 sectores industriales, reveló que el costo promedio de una fuga de datos en 2021 fue 4,24 millones de dólares, lo que representa un incremento del 10% en comparación con el año anterior (Tunggal, 2022).

En cuanto a proyecciones futuras, Cisco anticipó que para 2023 habría tres veces más de dispositivos conectados a la red que personas (Cisco, 2020). Por otro lado, Ventures proyecta que para 2025, el mundo almacenará 200 zettabaytes (2e14 GB) de datos, abarcando tanto infraestructuras públicas como privadas, como nubes y centros de datos, dispositivos personales e IoT (Morgan, 2010).

En años recientes, debido a los costos tanto económicos como no económicos que estos ataques internos malintencionados pueden acarrear, se ha prestado mayor atención al desafío de enfrentarlos. Como respuesta, se han desarrollado diversos métodos y técnicas para abordar este problema.

Las razones más destacadas para la implementación de mecanismos de prevención de pérdida de datos incluyen el cumplimiento de regulaciones y la protección de la propiedad intelectual (Forcepoint, 2020).

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); 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.; GDRP (General Data Protection Regulation, European Data Protection Regulation).

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

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.

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

- 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.

Kostadinov, en su artículo titulado Data Loss Protection (DLP) for ICS/SCADA, describe tres componentes esenciales que conforman la protección contra la pérdida de datos (Liu y Kuhn, 2010):

1. Identificar la información crítica.
2. Monitorear la transmisión de esa información.
3. Evitar el acceso no autorizado.

Además, DLP diferencia entre tres estados clave de los datos, cada uno requiriendo técnicas de protección específicas (Diario Oficial Unión Europea, 2016) (Securosis, L.L.C., 2014): Data-At-Rest (datos almacenados en dispositivos de almacenamiento); Data-In-Use (datos que están siendo utilizados por un usuario); Data-In-Motion (datos que se están transmitiendo a través de una red).

Entre las diversas tecnologías implementadas para proteger los datos en estos diferentes estados, se incluyen, entre otras, los Sistemas de Detección de Intrusiones (IDS) (Sans, 2017), los Sistemas de Prevención de Intrusiones (IPS), software antimalware, firewalls, actualizaciones de software y la Gestión de Eventos e Información de Seguridad (SIEM) (IBM, 2022) (Tahboud y Saleh, 2014).

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.

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. Ésta plataforma cuenta diversos tipos de tokens, entre ellos se puede mencionar Token DNS, Claves de AWS (notifica cuando alguien usa esas credenciales), Token log4shell (Hiesgen, Nawrocki, Schmidt, & Wählisch, 2022) (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.

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.

¿Qué pasa si algún documento o archivo con datos confidenciales es extraído fuera de su círculo de confidencialidad? El objetivo de este framework propuesto 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.

Desarrollo

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

En una primera fase, se propuso desarrollar un mecanismo que permita insertar un canary token en archivos (pdf, docx, xlsx, etc.), con el objetivo de obtener información sobre las circunstancias en las que se accedió al archivo. Esto permitiría determinar si dicho acceso puede considerarse una fuga de información, recolectar los datos pertinentes y activar las alertas necesarias.

Las preguntas que surgieron durante este proceso fueron:

- ¿Es indispensable tener trazabilidad para todos los documentos generados por una organización?
- ¿Cómo diferenciamos los documentos que requieren ser monitoreados de aquellos que no?
- ¿Qué documentos deben ser rastreados?
- ¿Qué información es esencial recopilar de cada documento que se está rastreando?

Para poder identificar el mecanismo que permita la inyección de canary tokens fue necesario investigar/estudiar 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.

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

- Un inyector de tokens a documentos (Fig. 1).
- Gestor del inyector de documentos (Fig. 1).
- Gestor de datos recibidos de documentos tokenizados.

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 Excel habilitado 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.

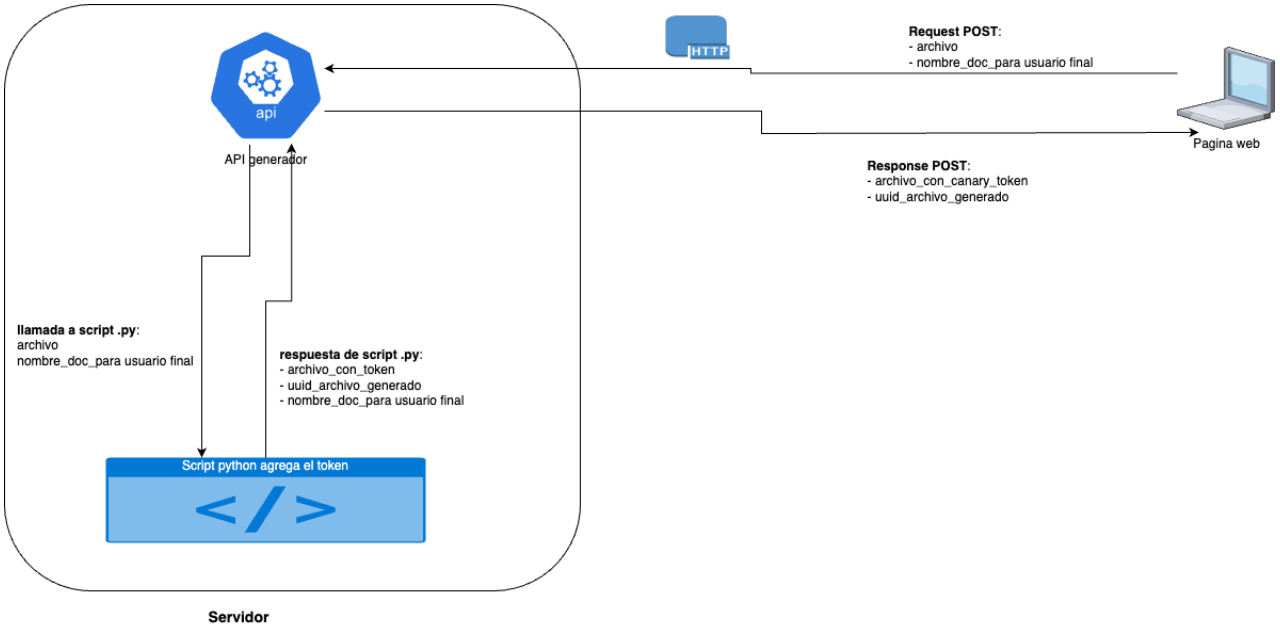

Fig. 1 Inyector de token a documentos

Por otro lado, se tiene el Visualizador de datos 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.

Conclusiones

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.

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.

Referencias

Arcserve. (2020). *The 2020 Data Attack Surface Report*. Arcserve Tape Backup Whitepaper. https://1c7fab3im83f5gqiow2qqs2k-wpengine.netdna-ssl.com/wp-content/uploads/2020/12/ArcserveDataReport2020.pdf (última visita: 15/08/2024).

Canary tokens. (n.d.). Página oficial. https://www.canarytokens.org/generate (última visita: 15/08/2024).

Código de Canary tokens. (n.d.). Github. Página oficial. https://github.com/thinkst/canarytokens (última visita: 15/08/2024).

ECMA International. (2021). *ECMA-376: Office Open XML file format* (5th ed.). https://ecma-international.org/publications-and-standards/standards/ecma-376 (última visita 15/08/2024).

Forcepoint. (n.d.). *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: 15/08/2024).

Hacktricks. (n.d.). *SSRF (Server Side Request Forgery)*. https://book.hacktricks.xyz/pentesting-web/ssrf-server-side-request-forgery (última visita: 15/08/2024).

Hiesgen, R., Nawrocki, M., Schmidt, T. C., & Wählisch, M. (2022). *The Race to the Vulnerable: Measuring the Log4j Shell Incident*. arXiv preprint arXiv:2205.02544.

Kostadinov, D. (2020). *Data Loss Protection (DLP) for ICS/SCADA*. https://resources.infosecinstitute.com/topic/data-loss-protection-dlp-for-ics-scada/ (última visita: 15/08/2024).

Morgan, S. (2020). *The World Will Store 200 Zetabytes of Data by 2025*. Cybersecurity Ventures. https://cybersecurityventures.com/cybersecurity-almanac-2020 (última visita: 15/08/2024).

National Institute of Standards and Technology (NIST). (n.d.). *Data Loss Prevention*. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=904672 (última visita: 15/08/2024).

Official Journal of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 (última visita 15/08/2024).

Papadimitriou, P., & Garcia-Molina, H. (2011). Data leakage detection. *IEEE Transactions on Knowledge and Data Engineering*, 23(1), 51-63. https://doi.org/10.1109/TKDE.2010.100

PCI Security Standards. (2022). *Payment Card Industry Data Security Standard: Requirements and Security Assessment Procedures* (v4.0). https://www.pcisecuritystandards.org/documents/PCI-DSS-v4_0-LA.pdf?agreement=true&time=1653751557059 (última visita: 15/08/2024).

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).

SANS Institute. (2017). *Intrusion Detection: SANS Reading Room*. https://www.sans.org/readingroom/whitepapers/detection/paper/38165 (última visita: 15/08/2024).

SANS Institute. (n.d.). *Understanding and Selecting a Data Loss Prevention Solution*. Securosis, L.L.C. https://securosis.com/assets/library/publications/DLP-Whitepaper.pdf (última visita: 15/08/2024).

Tahboub, R., & Saleh, Y. (2014). Data leakage/loss prevention systems (DLP). *International Journal of Information Systems*, 1(13), 13-19. https://doi.org/10.1109/WCCAIS.2014.6916624

Tunggal, A. (2022, mayo). *What is the Cost of a Data Breach in 2022?*. https://www.upguard.com/blog/cost-of-data-breach (última visita: 15/08/2024).

What is SIEM?. (n.d.). IBM. https://www.ibm.com/topics/siem (última visita: 15/08/2024).

The HIPAA Privacy Rule. (n.d.). https://www.hhs.gov/hipaa/for-professionals/privacy/index.html (última visita: 15/08/2024).