Canary token para la identificación temprana de fuga de datos

Canary token for early identification of data leaks

Presentación: 13 y 14 de setiembre de 2023

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

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

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

Leonardo Ciceri
Universidad Tecnológica Nacional – Facultad Regional Córdoba
leonardorciceri@gmail.com

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

Milagros Zea Cárdenas
Universidad Tecnológica Nacional – Facultad Regional Córdoba
milyzc@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

Ignacio Sánchez Balzaretti
Universidad Tecnológica Nacional – Facultad Regional Córdoba
ignaciojsb@gmail.com

Resumen

En los últimos años, considerando los costos económicos y no económicos que los ataques de fuga de datos 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 en una organización están la conformidad con regulaciones establecidas. 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. 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 clave: ciberseguridad, fuga de datos, Data Loss Prevention, prevención, Canary Token, seguimiento de activos digitales.

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 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.

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 (Yan and Kwon, 2014) (AFP, 2014) (Staff, 2016); 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 (Tahboub et al., 2014:13-19).

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) (Tunggal, 2022).

En cuanto a proyecciones para años futuros se puede mencionar:

- Según Cisco, para el 2023 se estimó que habría tres veces más dispositivos conectados a la red que humanos (Cisco, 2020).
- 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 (Morgan, 2020).

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 (Forcepoint, 2020).

Por su parte, Kostadinov en su artículo Data Loss Protection (DLP) for ICS/SCADA, explica los tres componentes fundamentales de DLP (Liu and Kuhn, 2010):

- 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 (Diario Oficial Unión Europea, 2016)(Securosis,L.L.C., 2014):

- 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) (Sans, 2017), Intrusion Prevention Systems (IPS), antimalwares, firewalls, actualizaciones de software y Security Information Event Management (SIEM) (IBM, 2022) (Tahboud and Saleh, 2014).

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.

Desarrollo

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 (Kumar AS., 2021):

- 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 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 (Digital Guardian, 2019).
- 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).

Actualmente, una de las plataformas más conocidas para generar canary tokens (Reale and Zinc, 2019: 66-68) (Fielding, et al., 1999) y de distintos tipos es canarytokens.org, creada por la organización Thinkst y de código abierto (Canary Tokens, 2021) (Github, 2022). 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 (R Hiesgen, 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.

Entre los hitos del proyecto para alcanzar los objetivos propuestos, se encuentran:

- 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 los 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?

Conclusiones

Actualmente, la plataforma de canary token (https://canarytokens.org/generate) 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 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.

Referencias

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.

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.

(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.

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

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

Cisco. (Febrero 2020) Cisco Annual Internet Report Forecasts 5G to Support More Than 10% of Global Mobile Connections by 2023.

The 2020 Data Attack Surface Report. Arcserve Tape Backup Whitepaper. Última visita: 10/05/2023.
https://1c7fab3im83f5gqiow2qqs2k-wpengine.netdna-ssl.com/wp-content/uploads/2020/12/ArcserveDataReport2020.pdf.

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.

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.

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

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.

SANS. (2017). SANS Institute: Reading Room - Intrusion Detection. https://www.sans.org/readingroom/whitepapers/detection/paper/38165.
What is SIEM?. https://www.ibm.com/topics/siem, última visita 15/04/2022].

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

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.

(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

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)

R. Fielding, J. Gettys, J. Mogul, H. Frysztyk, L. Masinter, P. Leach, T. Berners-Lee (1999). Hypertext Transfer Protocol -- HTTP/1.1. RFC-2616.
https://datatracker.ietf.org/doc/html/rfc2616.

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

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

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