DETCMAL
Plataforma de detección de anomalías en la red por medio de Machine Learning

Cardona R. Fredy, Cardozo M. Ximena,
Estudiantes de Maestría en seguridad de la información
Departamento de ingeniería de sistemas y computación
Universidad de los Andes, Colombia junio 2020

Resumen – La identificación temprana de actividad maliciosa o fuera de lo normal en una red de datos corporativa es una capacidad en creciente desarrollo que puede apoyar, el esfuerzo de identificación de eventos e incidentes relacionados con el compromiso de la confidencialidad, integridad o disponibilidad de la información valiosa, custodiada en los diferentes componentes de TI, mitigar impactos negativos a los negocios e incluso prevenir la materialización de riesgos. En este documento presentamos el proceso metodológico para construir un prototipo de análisis de tráfico basado de la identificación del comportamiento de dos tipos de amenazas conocidas y relevantes en el entorno del ciberespacio en Colombia, el cual se podrá utilizar para construir y ajustar un análisis de comportamiento en un entorno de red particular.

Índice de Términos – Análisis de comportamiento, Detección de anomalías, Machine learning y seguridad, Malware, Seguridad en Redes, Ramsoware.

I. CONTEXTO

En la actualidad, la mayoría de las organizaciones cuentan con sistemas informáticos para soportar sus negocios, procesando, almacenando y transfiriendo importantes volúmenes de información.

La información es un activo muy valioso y el robo o la filtración no autorizada de la misma puede causar grandes afectaciones, incluyendo el incumplimiento de regulaciones. Por su conexión a internet, hoy día las empresas deben considerar que siempre existe la posibilidad de que un atacante haciendo uso de la red afecte la operación crítica del negocio o extraiga información valiosa para este. Incluso, se han reportado casos, como Stuxnet, donde las empresas han sido atacadas aún sin tener una conexión directa a internet.[19]

A continuación, se muestran algunos datos estadísticos recopilados por Centro Cibernético Policial CECIP[1] sobre el caso colombiano:

- Colombia recibió el 30% de los ataques de Ransomware en Latinoamérica en el último año, seguido de Perú con el 16%, México 14%, Brasil 11% y Argentina 9%.
- La PYMES[2] fueron el blanco preferido por los atacantes, pues conocen que los niveles de seguridad son más bajos en este tipo de compañías.
- El crecimiento de los ataques de malware para Colombia fue 612% durante el último año[3].

Por otro lado, en el contexto regulatorio, las empresas del sector financiero, reguladas por la Super Intendencia Financiera a través de la circular 007 de 2018[4] son responsables de cumplir desde el 5 de junio de 2019, con los siguiente numerales:

4.2. Protección y detección

4.2.3. Realizar un monitoreo continuo a su plataforma tecnológica con el propósito de identificar comportamientos inusuales que puedan evidenciar ciberataques contra la entidad.

4.3. Respuesta y comunicación

4.3.1. Establecer procedimientos de respuesta a incidentes cibernéticos tales como: desconexión automática de equipos, cambios de contraseñas, actualizar la base de firmas del antivirus, bloqueo de direcciones IP o cualquier otro que determine la entidad.

Documento recibido el 25 de mayo de 2020. (Anote la fecha en que usted presentó su documento para su revisión.) Este trabajo fue apoyado en parte por el comité de proyectos del departamento de ingeniería de sistemas y computación de la Universidad de los Andes.

Cardona R, Fredy, estudiante de la maestría en seguridad de la información (e-mail: fa.cardona10@uninades.edu.co).

Cardozo M. Ximena, estudiante de la maestría en seguridad de la información (e-mail: ex.cardozo140@uniandes.edu.co)

Rueda R. Sandra, docente de la maestría en seguridad de la información (e-mail: sarueda@uniandes.edu.co)

[1] Tendencias del cibercrimen en Colombia, https://www.ccit.org.co/wp-content/uploads/informe-tendencias-final.pdf

[2] Pequeña y Mediana empresa

[3] En el contexto de informe entre el 2019 y 2020.

[4] Super Financiera de Colombia, normativa , Circular Externa 007 de 2018 - Imparte instrucciones relacionadas con los requerimientos mínimos para la gestión del riesgo de ciberseguridad. https://www.superfinanciera.gov.co/descargas/institucional/pubFile1031729/ance007_18.docx

Adicionalmente, en el uso de herramientas a nivel mundial, Gartner en la sección "Recomendaciones de mercado" escribe: "Las empresas deben considerar fuertemente el análisis NTA (*Network Traffic Analysis*) para complementar los métodos de detección basados en firmas y sandboxing. Muchos clientes de Gartner han informado que las herramientas NTA han detectado tráfico de red sospechoso que otras herramientas de seguridad perimetrales habían perdido"

Finalmente, MITRE, en su principio 4[5], iterar por diseño, resalta que el enfoque exitoso de la seguridad requiere una evolución iterativa y un constante refinamiento de modelos de seguridad, además de técnicas y herramientas para tener en cuenta el cambio de comportamiento adversario y comprender cómo las redes se ven comprometidas por un APT (Advance Persistent Threat).

II. DESCRIPCIÓN DE LA PROPUESTA

A. Descripción del problema

Las empresas hoy se enfrentan diariamente a ataques realizados por ciberdelincuentes o grupos organizados que buscan acceder a los sistemas informáticos para obtener información con la cual puedan lucrarse[6], ya sea cometiendo algún tipo de fraude, secuestrándola (ramsonware) o comercializándola en el mercado negro. Por este motivo es necesario que las organizaciones mantengan una postura de seguridad preventiva que permita hacer frente a este tipo de eventos realizando una detección oportuna de posibles eventos maliciosos en su red.

Entre los principales modus operandi de las amenazas cibernéticas encontramos:

- Los ciberdelincuentes utilizan ataques avanzados para evadir las soluciones de firewall, IPS y antivirus de última generación, y se ocultan en las organizaciones durante meses (320 días en promedio en 2015 al ser notificados de forma externa)[4].
- Más del 68 % del malware es exclusivo con respecto una organización y el 80 % de ese malware solo se utiliza una vez [5], lo que hace que las defensas basadas en firmas no sean eficaces contra los ataques específicos.
- Más del 80 % de las alertas por la seguridad basada en firmas y políticas no son confiables lo que evita que se enfoque en las alertas críticas [6].

Las organizaciones poseen o tienen acceso a muchas tecnologías de seguridad basadas en la red, que van desde firewalls de última generación hasta servidores proxy y entornos para pruebas de malware o sandboxing. Sin embargo, la efectividad de estas tecnologías se ve directamente afectada por su implementación. Demasiada confianza en las capacidades integradas como control de aplicaciones, antivirus, prevención de intrusiones, prevención de pérdida de datos u otros motores automáticos de detección profunda de paquetes de detección del malware conduce a una implementación altamente preventiva, con grandes brechas tanto en prevención como en detección.

Nuestra solución, DETCMAL (Detección de malware), se centra en el uso de soluciones de seguridad de capa de aplicación que una organización ya posee con una mentalidad moderna, al diseñar defensas para ataques sofisticados como los APT (Advanced Persistent Threats), las capacidades de prevención y detección aumentan significativamente.

Monitoreo de seguridad de la red: las alertas de detección de intrusiones y los metadatos de la red brindan un enfoque holístico para obtener una línea base de comportamiento e identificar actividades no autorizadas. Esta solución se enfoca en detectar malware que opera a través de la red con NSM (Network Security Monitoring). Esta solución al utilizar machine learning mejora la efectividad en la detección de comportamientos no conocidos en comparación con las técnicas que detectan con base en firmas.

B. Propuesta de solución

Hoy existen diversos modelos de analítica de datos que permiten realizar la detección de anomalías o comportamientos maliciosos en la red con cierto grado de efectividad. Lo que busca este trabajo es evaluar la efectividad de un conjunto de modelos en un contexto determinado, identificar el más efectivo y proponer una aproximación, basada en inteligencia artificial, que permita incrementar su efectividad, apoyando de esta manera a las organizaciones en su capacidad de detección y respuesta oportuna ante eventos de seguridad.

C. Objetivos.

Identificar la precisión con la que un modelo de representación de datos, de un conjunto de modelos seleccionado, detecta variaciones de tipo malicioso o anómalo en una red de datos corporativa de una pequeña o mediana empresa. El trabajo analiza detalladamente malware del tipo troyano como Emotet y sus variantes.

A continuación se presentan los objetivos específicos asociados:

- Seleccionar los cuatro modelos que serán parte de la evaluación.

[5] MITRE, technical report. Finding Cyber Threaths, with attacks – Based analytics https://www.mitre.org/sites/default/files/publications/16-3713-finding-cyber-threats%20with%20att%26ck-based-analytics.pdf

[6] Tendencias del cibercrimen en Colombia, https://www.ccit.org.co/wp-content/uploads/informe-tendencias-final.pdf. "El principal interés de los Cibercriminales en Colombia se basa en la motivación económica y la posterior monetización de las ganancias generadas en cada Ciberataque."

- Diseñar las pruebas que permitan identificar el modelo que mejor se ajuste al contexto de la organización.
- Definir e implementar la arquitectura de recolección, procesamiento y almacenamiento de los datos de pruebas.
- Documentación de parámetros de aceptación del modelo.

D. Propuesta de solución

La solución propuesta consta de las siguientes capas:

- Capa de recolección de datos
- Capa de transporte de datos
- Capa de procesamiento de datos
- Capa de almacenamiento de datos
- Capa de presentación de resultados

1) Capa de recolección de datos: En esta capa se utilizan diferentes fuentes para recabar datos del tráfico de red, como flujos de PCAP, eventos generados por SNORT, entre otros.

Para recolectar los datos de tráfico necesarios para crear una línea base se propone una solución basada en hardware que a través de un tap en la red recopila tráfico durante un tiempo de 20 a 30 días. Si la recolección de tráfico no es posible, se puede hacer uso de data sets recolectados previamente, o utilizar una solución de software basada en agentes.

2) Capa de transporte: para el manejo del stream de datos. Esta corresponde a la necesidad de contar con un bus que permita manejar todas las fuentes de datos que se generan desde la capa de los colectores.

3) Capa de procesamiento de datos: Así como su nombre lo indica en esta capa se ejecutan los diferentes procesos de enriquecimiento y análisis de datos.

En esta capa se implementará el modelo más efectivo para la detección de malware tipo troyano como Emotet. Los principales criterios de selección de los modelos son:

- Capacidad para determinar categorías
- Manejo de datos etiquetados
- Manejo de volúmenes de datos mayores a 100.000 muestras
- Asertividad (assertiveness) del modelo

Las principales variables que se utilizaran en el entrenamiento del modelo serán:

- Enumeración de todos los equipos y dispositivos que tengan recursos compartidos
- Acceso de nuevas MAC a la Wifi con contraseñas débiles
- Equipos con SMB habilitado y mayor volumen de tráfico o interacciones en la red

4) Capa de almacenamiento. Solución Big Data para almacenamiento de grandes volúmenes de información.

5) Capa de presentación. Tablero general con acceso a reportes e indicadores para la gestión respectiva.

E. Implementación del prototipo

El primer paso para la implementación del prototipo fue la selección de los tres algoritmos de ML que serían apropiados para el objetivo, es decir, para identificar tráfico anómalo.

Para determinar los algoritmos seguimos el árbol de decisión presentado en la Fig1.

En nuestro caso utilizamos algunos sets de datos públicos de más de 100000 muestras con etiquetas que identifican el tráfico normal y el anómalo lo que permite utilizar algoritmos clasificadores.

Para el prototipo se hizo uso de algoritmos de árboles de decisión y bayesianos.

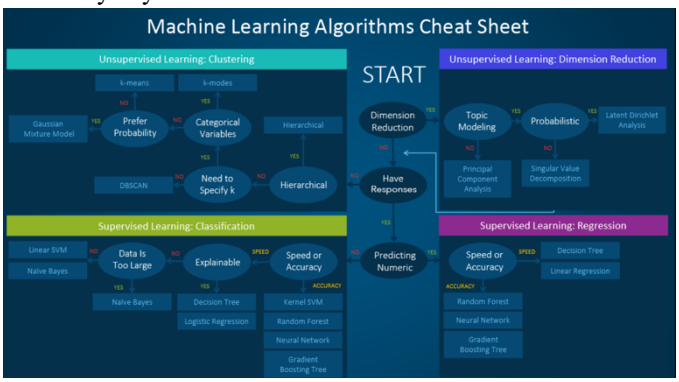


Fig. 1. Selección de los algoritmos de ML apropiados para el logro del objetivo. Fuente[7]

Para la construcción de los modelos se utilizó la herramienta Weka, una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato de Nueva Zelanda. Weka es software libre distribuido bajo la licencia GNU-GPL. Esta herramienta permite de manera gráfica aplicar diferentes algoritmos de clasificación de manera sencilla e intuitiva.

De los conjuntos de datos se tomaron como variables de interés los protocolos, puertos, direcciones IP tanto origen como destino. Los resultados obtenidos aplicando el algoritmo J48 de árboles de decisión, Random Tree y NaiveBayesian se muestran a continuación:

ALGORITMO	CLASIFICADAS CORRECTAMENTE	CLASIFICADAS ERRONEAMENTE
J48	92.0 %	7.9 %
RandomTree	51.4 %	48.5 %

[7] https://blogs.sas.com/content/subconsciousmusings/files/2017/04/machine-learning-cheet-sheet-2.png.

NaiveBayes	92.3 %	7.6 %

Tabla 1. Resultados de clasificadores utilizados

El conjunto de datos para el entrenamiento y pruebas tiene 6000 muestras, que se repartieron en 66% para entrenamiento y 34% para comprobación.

De los tres modelos probados el que arrojó mejor porcentaje de acierto (asertivity) corresponde con el NaiveBayes, dato que por sí solo no es confiable dado que el porcentaje de datos de tipo anómalo es muy bajo.

	Tráfico normal		Tráfico malicioso	
Algoritmos	**TP Rate**	**FP Rate**	**TP Rate**	**FP Rate**
NaivesBayesian	0.918	0.000	1.000	0.082
J48 (DecisionTree)	1.000	0.457	0.543	0.000
RandomTree	0.999	1.000	0.000	0.001

Tabla 2. Comparativo resultados

Del resultado queda como posible continuación de este trabajo la evaluación de nuevas variables que permitan mejorar la asertividad del(os) modelo(s) y de la aplicación de algoritmos de Deep Learning.

III. PROTOTIPO

El prototipo se ejecutará con herramientas Open Source y herramientas administradas de nube para habilitar el pago por demanda:

1. Recolección de datos, amazon kinesis data firehose
2. Análisis de datos, amazon kinesis Data analytics
3. Procesamiento de datos, Amazon reddshit o Amazon athena.
4. Visualización, Amazon quicksight o kibana "Business Intelligence" y "Monitoring" herramientas de monitoreo respectivamente

Configurará un agente de Kinesis en orígenes de datos para recopilar datos y enviarlos de manera continua a Amazon Kinesis Firehose.

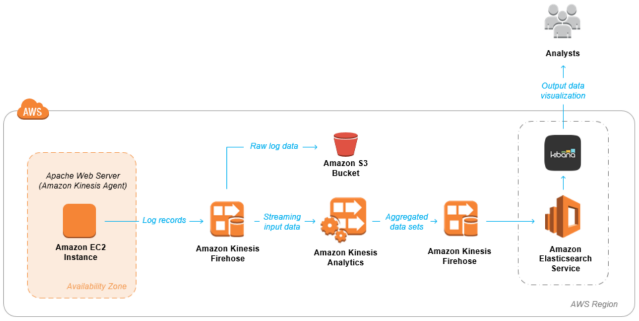
Creará un flujo de entrega de datos integral con Kinesis Firehose. El flujo de datos transmitirá los datos desde el agente a los destinos, incluidos Amazon Kinesis Analytics, Amazon Redshift, Amazon Elasticsearch Service y Amazon S3.

Procesará datos de registro entrantes mediante consultas SQL en Amazon Kinesis Analytics.

Cargará datos procesados de Kinesis Analytics en Amazon Elasticsearch Service para indexarlos.

Analizará y visualizará los datos procesados con Kibana.

A futuro, la solución podrá incorporar datos de streaming con Kinesis Data Streams, procesarlos con Kinesis Data Analytics y enviar los resultados a cualquier aplicación o almacén de datos mediante Kinesis Data Streams con una latencia integral en milisegundos. Esto implica una solución en “tiempo real”.



Para el desarrollo del prototipo, se utilizaron data sets, lo cual elimina la necesidad de usar amazon kinesis data firehose, por tanto, la propuesta de arquitectura, en tiempo “no real”, sería la siguiente:

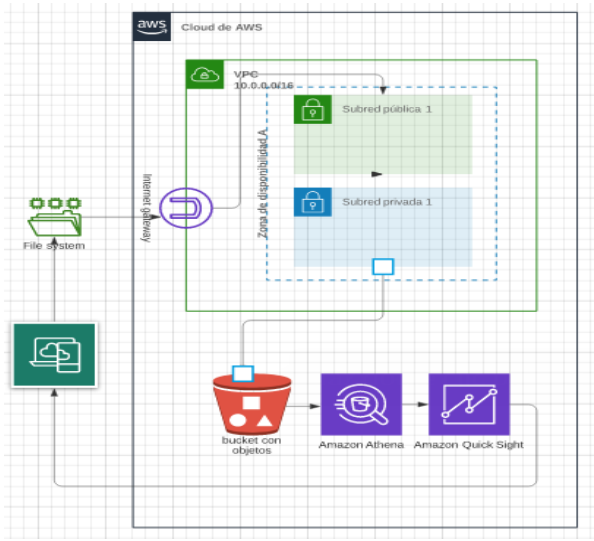


Fig. 2. Arquitectura de prototipo implementada. Figura creada por los autores.

Para el alcance de este trabajo final se aplicará una arquitectura más simple enfocada en validar la viabilidad de la solución.

IV. ESTRATEGIA DE NEGOCIO

La estrategia de publicación de la solución para que la comunidad tenga acceso tiene dos componentes:

Publicación del software. Documentar el diseño e implementación y publicarlo en GitHub, para que el público general tenga acceso, mejore la idea y aumente su alcance.

Servicios. El segundo componente es la venta de servicios de apoyo a la implementación y afinamiento de la herramienta. La venta de servicios profesionales asociados, pueden convertirse en un modelo de negocio. La siguiente sección presenta los costos asociados con este caso.

A. Manejo de costos y presupuestos

En la medida que se utilicen servicios administrados, solo se pagará por el flujo de tráfico y el almacenamiento, en costes, que van desde 51 a 100 dólares mensuales.

Mientras tanto, en la siguiente tabla se resumen los costos asociados a la implementación y puesta en marcha del prototipo:

Costos por etapas:									
	Costo Consultor seguridad de la información			Costo Data Analyst			Costos Infraestructura		
Etapa	Horas	Valor Hora	Valor Total	Horas	Valor Hora	Valor total	Servidor	Almacenamiento	Total
Entendimiento del contexto	240	$ 102.000	$ 24.480.000	40	$ 85.000	$ 3.400.000	$ 0	$ 0	$ 27.880.000
Evaluación estratégica	0	$ 102.000	$ 0	160	$ 85.000	$ 13.600.000	$ 1.496.000	$ 147.348	$ 15.243.348
Despliegue de capacidades de identificación en la red.	160	$ 102.000	$ 16.320.000	0	$ 85.000	$ 0	$ 475.600	$ 305.122	$ 17.100.722
Elaboración del informe de conclusiones	120	$ 102.000	$ 12.240.000	60	$ 85.000	$ 5.100.000	$ 475.600	$ 419.922	$ 18.235.522

Tabla. 1. Costos asociados al diseño e implementación del prototipo.

Cambien el formato de la tabla (a lo largo porque el tamaño de la letra en este formato no es razonable).

IX. CONCLUSIONES

Machine Learning es una alternativa viable para la detección de eventos e incidentes de seguridad como ramsomware y malware:

- Disminuye el tiempo de identificación de incidentes de seguridad por medio de su automatización en el manejo y análisis de grandes volúmenes de información.
- Para el caso de ransomware y malware es necesario contar con un data sets reciente y con un gran número de registros, con el fin de mejorar el umbral de detección.
- Para difundir el uso del producto, resulta favorable el ofrecer los scripts de despliegue en Cloud, pues disminuye los costos de implementación y operación.
- EL producto está enfocado para el uso de profesionales responsables de administrar la red de datos y operadores de Centros Operativos de seguridad SOC
- Es importante indicar que cualquier mdoelo generado por medio de machine learning esta limitado a la calidad de los datos y sus condiciones, que pueden producir sesgos en los modelos o que se han “ciegos” a cierto tipo de eventos únicos dentro del conjunto de datos.

IX. TRABAJO FUTURO

Consideramos que la riqueza de este tipo de trabajos de grado proviene de la participación interdisciplinaria, si este trabajo, se toma como base, para un trabajo de maestría de ingería de datos, es posible que se obtenga una versión refinada, de la recolección, transformación y análisis de los datos. La perspectiva de los ingenieros que desarrollamos este proyecto es principalmente, de seguridad de la información.

RECONOCIMIENTO

Agradecimientos de los autores a la Universidad de los Andes y a los docentes Harold Enrique Castro Barrera; Ricardo Gómez Diaz; Diego Hernán Pérez Jaramillo; Sandra Julieta Rueda Rodríguez; .por su apoyo en la definición, acompañamiento y soporte en la consecución de este proyecto.

REFERENCIAS

[1]https://www.gartner.com/en/documents/3945589/market-guide-for-intrusion-detection-and-prevention-syst
[2]https://www.gartner.com/en/documents/3945589/market-guide-for-intrusion-detection-and- prevention-syst
[3]https://www.gartner.com/en/documents/3902353/market-guide-for-network-traffic-analysis
[4] FireEye (febrero de 2016). M-Trends 2016.
[5] Joshua Goldfarb (19 de septiembre de 2016). “Detection Innovations.”
[6] 3 Ponemon Institute LLC (enero de 2015). “The Cost of Malware Containment.”
[7]https://www.cisco.com/c/dam/en/us/products/collateral/security/small-mighty-threat.pdf
[8] https://security-report.eset-la.com/
[9] Informe de las tendencias de cibercrimen en Colombia (2019-2020)
[10] RISTI, N.º E27, 03/2020 - https://www.researchgate.net/journal/1646-9895_RISTI-Revista_Iberica_de_Sistemas_e_Tecnologias_de_Informacao
[11]https://www.elastic.co/es/blog/analyzing-network-packets-with-wireshark-elasticsearch-and-kibana
[12] Baesesns, Vlasselaer, Verbeke. Fraud Analytics. Using Descriptive, predictive, and social network techniques.
[13] Getting Satrted with WEKA. Machine learning recipes https://www.cs.waikato.ac.nz/ml/weka/
[14] Alpaydin, Ethem. Introduction to Machine Learning. 2014
[15] Baxter, James H. Wireshark Essentials. 2016
[16] Tanya Garg. Comparison of Classification Techniques for Intrusion Detection Dataset Using WEK. 2014
[17]https://www.microsoft.com/security/blog/2018/02/14/how-artificial-intelligence-stopped-an-emotet-outbreak/
[18] Ravanshad Abolfazl . How to choose machine learning algorithms?. https://medium.com/@aravanshad/how-to-choose-machine-learning-algorithms-9a92a448e0df
[19] https://www.sans.org/cyber-security-summit/archives/file/summit-archive-1493844778.pdf
[20] https://mcfp.felk.cvut.cz/publicDatasets/CTU-Mixed-Capture-5/