Análisis y detección de amenazas de los posibles ataques de tipo Comando y Control en el COCIB mediante métodos combinados de Ciencias de Datos e Inteligencia Artificial

Primera Fase

8 de Julio de 2024
UNIVERSIDAD ICESI
Maestria en ciencias de datos

Análisis y detección de amenazas de los posibles ataques de tipo Comando y Control en el COCIB mediante métodos combinados de Ciencias de Datos e Inteligencia Artificial

Integrantes

Javier Ricardo Muñoz

Yesid Humberto Montaño

Laura Espinosa

Andrea Timaran

Carlos Enrique Jaramillo

Yasmin Johanna Garcia

Tutor

PhD. Uram Anibal Sosa

Análisis y detección de amenazas de ~~los posibles~~ ataques de tipo Comando y Control en el COCIB mediante métodos combinados de Ciencias de Datos e Inteligencia Artificial

Armada Nacional de Colombia

Comando de cibernética Naval

2024

Contenido

Introducción

El ámbito de la ciberseguridad ha experimentado un incremento notable en la frecuencia y sofisticación de los ciberataques en los últimos años, sobre todo aquellos destinados a infraestructuras críticas. El aumento no solo ha afectado a sectores como la energía, la salud y las finanzas, sino que también ha convertido a entidades militares y de defensa en objetivos prioritarios para actores maliciosos. En este contexto, el ~~Centro Operativo de Cibernética Naval de la Armada de Colombia (COCIB) se ha~~ transformado en un objetivo constante, representando una amenaza creciente para la seguridad cibernética y la integridad de las operaciones de las Fuerzas Armadas ~~Navales de Colombia,~~ y, por lo tanto, para la seguridad nacional.

La infraestructura del ~~COCIB~~, un componente estratégico para la defensa y el funcionamiento efectivo de la Armada, se enfrenta a desafíos constantes en términos de detección, prevención y mitigación de estos ataques. La necesidad de abordar este problema de manera integral y proactiva se ve agravada por la sofisticación y frecuencia crecientes de los ciberataques. La seguridad de la infraestructura ~~del COCIB~~ no solo es vital para proteger los activos y datos críticos de la Armada, sino también esencial para garantizar la efectividad de las operaciones navales y la seguridad nacional en general.

En este contexto, es importante comprender y abordar ~~estos tipos específicos~~ de amenazas cibernéticas considerando la naturaleza crítica del COCIB y el impacto potencial de ataques exitosos de Comando y Control. Estos ataques, que permiten a los cibercriminales mantener el control de sistemas comprometidos ~~y filtrar información de manera encubierta,~~ pueden tener consecuencias devastadoras si no se detectan y mitigan a tiempo. Por lo tanto, es necesario desarrollar estrategias efectivas de detección, prevención y respuesta ante estas amenazas.

La integración de técnicas avanzadas de Ciencia de Datos e Inteligencia Artificial se presenta como una solución prometedora para abordar estos desafíos. Estas tecnologías permiten el análisis de grandes volúmenes de datos, la detección de anomalías, patrones y la predicción de comportamientos maliciosos que podrían pasar desapercibidos con métodos tradicionales. Los algoritmos de aprendizaje automático y otras técnicas de IA mejoran la capacidad de detección y permiten una respuesta más proactiva y eficiente a posibles ataques.

Al automatizar tareas de monitoreo y análisis, se liberan recursos ~~humanos,~~ permitiendo que el personal se concentre en actividades estratégicas, como la elaboración de políticas de seguridad y la planificación de medidas preventivas. Implementar estos métodos en los datos provenientes del Rapid7, activo importante del COCIB, es esencial para fortalecer la postura de ciberseguridad de la

Armada de Colombia y garantizar la protección de sus operaciones críticas. La adopción de estas tecnologías avanzadas no solo representa una evolución en las capacidades defensivas del COCIB, sino que también sienta un precedente para la modernización de la ciberseguridad en otras entidades críticas del país.

Palabras claves

Amenaza, Ataque, Ciberseguridad, Ciberespacio, Cibernética Naval, Seguridad informática, Vulnerabilidad, COCIB, Ciencia de Datos, Inteligencia Artificial, Machine Learning, Rapid7, Patrones y anomalías, Comando y Control (C&C).

Identificación del Problema

~~La vulnerabilidad del Centro Operacional Cibernético Naval de Colombia (COCIB) f~~rente a la creciente sofisticación y frecuencia de los ciberataques.

En los últimos años, las Fuerzas Armadas a nivel global han experimentado un aumento significativo en la cantidad y complejidad de los ciberataques contra su infraestructura crítica. Esta tendencia está causando una preocupación por el impacto potencial de estos ataques en la ciberseguridad y la integridad de las operaciones militares, así como sus implicaciones directas para la seguridad nacional. La infraestructura ~~del Centro Operacional Cibernético Naval de Colombia (COCIB),~~ como componente estratégico para la defensa y la operación efectiva de la Armada, enfrenta constantes desafíos en la detección, prevención y mitigación de ciberataques, que van desde pequeñas intrusiones hasta ataques sofisticados de Comando y Control (C&C). La creciente sofisticación y frecuencia de estas ciberamenazas plantean desafíos importantes, lo que pone de relieve la necesidad urgente de abordar este problema de manera integral para permitir una respuesta más rápida y eficaz a los incidentes de seguridad.

La seguridad de las infraestructuras del COCIB es vital no solo para proteger los activos y datos críticos de la Armada, sino también para garantizar la eficiencia y continuidad de las operaciones navales y, en última instancia, la seguridad nacional en su conjunto. Por tanto, es necesario estudiar en profundidad los factores que contribuyen al aumento de los ciberataques dirigidos contra el ~~COCIB,~~ así como desarrollar estrategias efectivas de detección, prevención y respuesta frente a estas amenazas.

Comprender la naturaleza y el alcance de estos ataques es esencial para implementar medidas de protección adecuadas y fortalecer la postura de ciberseguridad de la Armada de Colombia. En particular, es fundamental que estas acciones sean proactivas y no reactivas, anticipando los riesgos y minimizando su impacto antes de que se comprometa la seguridad y las operaciones navales. La

implementación de estrategias avanzadas basadas en ciencia de datos e inteligencia artificial es crucial para mejorar las capacidades de detección ~~y~~ respuesta, asegurando así la resiliencia y operatividad del COCIB y, en consecuencia, ~~de las Fuerzas Armadas Navales de Colombia.~~

En este sentido, las preguntas que guían este proyecto son:

- ¿Cuáles son los patrones de comportamiento malicioso más comunes en los registros (logs) de dispositivos críticos ~~del COCIB~~ asociados a ataques de Comando y Control (C&C)?
- ¿Qué modelos de Ciencia de Datos e Inteligencia Artificial son más efectivos para detectar y caracterizar estos patrones de comportamiento malicioso en tiempo real?
- ¿Cómo se puede implementar un sistema de detección de ataques C&C basado en Ciencia de Datos e Inteligencia Artificial que sea escalable, adaptable y capaz de integrarse con la infraestructura existente del COCIB?

Justificación

En un entorno donde las amenazas cibernéticas dirigidas a infraestructuras críticas se multiplican y se vuelven cada vez más sofisticadas, se subraya la necesidad urgente de fortalecer las capacidades de ciberseguridad ~~del Centro Operativo de Cibernética Naval de~~ la Armada de Colombia **(**COCIB). La seguridad de esta infraestructura es esencial no solo para proteger los activos y datos críticos de la Armada, sino también para garantizar la efectividad y continuidad de sus operaciones navales y, en última instancia, la seguridad nacional en su conjunto. Por lo tanto, es crucial desarrollar estrategias efectivas para identificar y mitigar posibles ~~vulnerabilidades q~~ue podrían convertirse en ataques.

El crecimiento exponencial de las amenazas cibernéticas plantea un desafío sin precedentes para la seguridad informática en todas las industrias, especialmente en aquellas relacionadas con la seguridad nacional. Los ataques, cada vez más sofisticados y difíciles de detectar, ponen en riesgo la integridad y confidencialidad de la información crítica, así como la continuidad operativa de las organizaciones, y en particular, de las naciones.

Los ataques de tipo Comando y Control representan una amenaza particularmente grave, ya que permiten a los atacantes establecer canales de comunicación encubiertos dentro de las redes comprometidas, facilitando el robo de información, el sabotaje de sistemas y la coordinación de actividades maliciosas. Para hacer frente a estas amenazas, es imperativo adoptar un enfoque proactivo que incluya la identificación y caracterización de patrones de comportamiento malicioso en los registros de dispositivos críticos.

La aplicación de técnicas avanzadas de Ciencia de Datos se presenta como una solución prometedora para mejorar la detección temprana de ciber amenazas. Los métodos de Ciencia de Datos ofrecen una capacidad robusta para analizar grandes volúmenes de datos de forma rápida y precisa. Mediante el uso de algoritmos de Machine Learning y otras técnicas derivadas de la Inteligencia Artificial, es posible identificar patrones y anomalías que podrían pasar desapercibidos para los enfoques tradicionales de detección, que por lo general dependen de la capacidad humana. Esta capacidad predictiva y analítica se convierte en un recurso invaluable para anticipar y responder eficazmente a las amenazas cibernéticas emergentes.

Además, la integración de métodos de Ciencia de Datos en la infraestructura de ciberseguridad existente no sólo mejora la capacidad de detección, sino que también permite una respuesta más rápida y eficiente ante posibles ataques. Al automatizar tareas de monitoreo y análisis, los sistemas basados en Ciencia de Datos liberan recursos humanos, permitiendo que el personal se concentre en actividades de mayor valor estratégico, como la elaboración de políticas de seguridad y la planificación de medidas preventivas.

En este sentido, el presente trabajo no solo abordará los desafíos actuales de ciberseguridad que enfrenta el COCIB, sino que también contribuirá a una defensa más sólida y resiliente de ~~las Fuerzas Armadas Navales de Colombia~~.

Objetivos

Objetivo general

Implementar métodos avanzados de Ciencia de Datos e Inteligencia Artificial, con un enfoque proactivo para proteger los activos y operaciones esenciales para fortalecer la ciberseguridad del ~~Centro Operativo de Cibernética Naval~~ de la Armada Nacional de Colombia (COCIB) mediante la identificación y caracterización de los patrones de comportamiento malicioso en los registros (logs) de dispositivos críticos frente a ataques de tipo Comando y Control, utilizando.

Objetivos específicos

- Elaborar un Análisis Exploratorio de Datos (EDA) de los conjuntos de información proporcionados por la Fuerza Naval de Colombia, con un enfoque en los sistemas de Comando y Control, para comprender la naturaleza y distribución de los datos, identificando valores atípicos (outliers), patrones recurrentes y anomalías relevantes, y así caracterizar posibles amenazas cibernéticas.

- Diseñar y entrenar diversos modelos de clasificación y regresión utilizando algoritmos de aprendizaje supervisado como árboles de decisión y SVM, así como algoritmos no supervisados como Clustering, para evaluar su desempeño en la detección de amenazas cibernéticas de tipo Comando y Control. Emplear métricas como precisión, recall y F1-score para evaluar y comparar la efectividad de los modelos desarrollados.
- Validar y desplegar los modelos seleccionados en un entorno operativo simulado, utilizando conjuntos de datos independientes y técnicas de validación cruzada para garantizar su generalización y robustez. Los modelos serán implementados en una infraestructura adaptada para integrarse con los sistemas de ciberseguridad existentes de la Fuerza Naval de Colombia, asegurando su eficacia y escalabilidad en la detección y análisis en tiempo real de eventos potencialmente peligrosos.

Alcances 2024-1

Durante esta primera fase se hace fundamental establecer los alcances que se planean para el periodo 2024-1. Estos alcances se ven limitados por tiempo, disposición y entendimiento del negocio, análisis de datos y exploración de herramientas.

En este sentido los alcances para este periodo son:

Objetivo terminal 1: Entendimiento del negocio

Comprender el contexto operativo, los objetivos estratégicos y las necesidades del COCIB en relación con los ciberataques de tipo comando y control, analizando los procesos, flujos de trabajo, actores involucrados y principales amenazas cibernéticas.

Objetivo terminal 2: Criterios de limpieza, imputación y selección de variables

Seleccionar y aplicar criterios para la limpieza de datos, imputación de valores faltantes y selección de variables relevantes. Esto incluye identificar y corregir errores en los datos, determinar los métodos más adecuados para imputar valores faltantes y escoger las variables que cumplan estos criterios.

Objetivo terminal 3: Detección de anomalías más comunes en los registros

Identificar anomalías en los registros de datos utilizando técnicas de visualización. Esto incluye la detección de valores atípicos y patrones inusuales, así como la contextualización de estas anomalías para determinar su relevancia en términos de amenazas cibernéticas.

Marco Teórico

El incremento y la sofisticación de los ataques en el ciberespacio han puesto en alerta a las agendas nacionales, destacando la necesidad urgente de fortalecer las defensas cibernéticas. La Agencia Europea de Ciberseguridad (ENISA) ha revelado un análisis exhaustivo de más de 2.500 ciberataques mayores entre julio de 2022 y junio de 2023, donde el 19% se dirigieron contra administraciones públicas, subrayando la vulnerabilidad en este sector. Este aumento en la sofisticación de las herramientas disponibles para perpetrar ciberataques se acompaña de una creciente tensión geopolítica, mientras que la continua innovación en la industria digital plantea nuevos desafíos para la ciberseguridad. Además, la evolución tecnológica exige el desarrollo de nuevos marcos regulatorios para garantizar una protección adecuada, desde la fijación de reglas para la implementación de Inteligencia Artificial hasta el establecimiento de estándares de ciberseguridad.

Se observa, en general, que tanto en el ámbito corporativo como en el sector público existen algunas insuficiencias en cuanto a la cultura de la ciberseguridad. Esto evidencia una falta generalizada de conciencia y participación en las prácticas recomendadas para protegerse en el entorno digital, así como la carencia de protocolos adecuados para enfrentar los incidentes y recuperarse de los mismos, especialmente preocupante en entornos con alta rotación de personal. Estos desafíos, considerados críticos, incluyen la dificultad para cubrir los perfiles necesarios para la gestión de la ciberseguridad, la concentración del poder digital, la escasa implementación de estándares internacionales, la ausencia de una estructura nacional de gobierno ejecutivo, la falta de arquitectura resiliente en las organizaciones y la dificultad para perseguir a los ciberdelincuentes. Ante este panorama, es esencial que tanto las organizaciones corporativas como las entidades del sector público aumenten su atención y dedicación a la gestión de estos riesgos para garantizar su seguridad digital y proteger sus activos críticos.

En este contexto, el COCIB ha reconocido la urgencia de implementar modelos de Ciencia de Datos e Inteligencia Artificial centrados en la detección de amenazas cibernéticas. Estos modelos tienen como objetivo mejorar la predicción, análisis, procesamiento, visualización y extracción de características de eventos potencialmente peligrosos, utilizando registros provenientes de diversas herramientas de monitoreo cibernético y de inteligencia de amenazas. Este macroproyecto se dividirá en varios subproyectos interconectados que abordarán diferentes aspectos de la ciberseguridad. Se aprovecharán las capacidades del aprendizaje automático y del aprendizaje profundo para generar conocimientos más precisos y efectivos en la monitorización de eventos que representen un riesgo real.

Ciberseguridad

La ciberseguridad abarca el conjunto de medidas, tecnologías y prácticas diseñadas para proteger los sistemas, redes y datos de ataques maliciosos en el ciberespacio. Esta busca garantizar la confidencialidad, integridad y disponibilidad de la información, salvaguardando tanto a individuos como a organizaciones de amenazas como el robo y secuestro de datos, el fraude, el espionaje y la denegación de servicios. Si bien no se tiene un momento exacto donde se haya hablado de ciberseguridad por primera vez, Thoma Creeper (1975) marcó el inicio de la preocupación por la seguridad en los sistemas informático, si bien diversos autores han marcado internacionalmente las pautas de la ciberseguridad informática, algunos autores destacados en este campo incluyen a Bruce Schneier, reconocido por su enfoque en la criptografía y la seguridad informática, y Dan Kaminsky, conocido por sus importantes contribuciones en la detección y resolución de vulnerabilidades en internet, experto en seguridad de redes y descubridor de vulnerabilidades críticas.

Análisis de la ciberseguridad en Colombia

Colombia enfrenta desafíos significativos en materia de ciberseguridad, como lo evidencian los numerosos ataques cibernéticos reportados en los últimos años. Según la revista Forbes, para el año 2024, Colombia ha ocupado por segunda vez el segundo puesto en ser el país de América latina con más ciberataques[1]. La falta de conciencia y capacitación en seguridad digital, junto con la rápida adopción de tecnologías, aumentan la vulnerabilidad del país. Sin embargo, se han realizado avances importantes, como la creación del Centro de Respuesta a Incidentes Cibernéticos de Colombia (ColCERT) y la implementación de políticas y estrategias para fortalecer la ciberseguridad a nivel nacional.

En este sentido, este proyecto se enfoca en fortalecer la ciberseguridad del Comando de Cibernética Naval, desarrollando estrategias y herramientas para prevenir y mitigar los ataques cibernéticos más comunes que amenazan la integridad de sus sistemas y datos. Se centrará en identificar vulnerabilidades, implementar medidas de protección y establecer protocolos de respuesta ante incidentes para garantizar la seguridad de la información crítica y la continuidad de las operaciones navales.

1.[1]Forbes. (2024). *Colombia es el país con más ataques de ciberseguridad en Latinoamérica*. Recuperado de https://forbes.co/2024/02/28/tecnologia/colombia-es-el-pais-con-mas-ataques-de-ciberseguridad-en-latinoamerica Recuperado el 7 de junio de 2024

Marco legal

La Armada Nacional de la República de Colombia (ARC) hace parte de una de las cuatro fuerzas militares del país, su principal objetivo es contribuir y garantizar la defensa del Estado y las zonas marítimas de su jurisdicción mediante la aplicación de su poder naval.[2] Es la fuerza que deficiente los ríos, lagos, mares y océanos de Colombia.

La ARC se extiende a lo largo del territorio colombiano mediante una red de Bases Navales (BN) estratégicamente distribuidas. Esta red incluye un número de bases navales y fluviales en ambos litorales del país, además de múltiples bases fluviales que se encuentran ubicadas estratégicamente a lo largo del territorio nacional.

Las Bases Navales mayores son:

- Base Naval ARC Bolívar -BN1, cerca de Cartagena
- Base Naval ARC Bahía Málaga - BN2, cerca de Buenaventura
- Base Naval ARC Leguizamo - BN3, cerca de Puerto Leguizamo
- Base Naval ARC San Andrés - BN4, en San Andrés
- Base Naval ARC Orinoquía - BN5, en Puerto Carreño
- Base Naval ARC Bogotá - BN6, en Bogotá

El Comando de cibernética naval

El comando de cibernética naval o COCIB, recibe su denominación, sigla y estructura en la resolución 127 del año 2021[3]:

Artículo 165° Cambiar denominación, sigla y modificar la estructura interna de la Dirección de Cibernética Naval, sigla (DICIB), por Comando de Cibernética Naval, sigla (COCIB), orgánico de la Jefatura de Inteligencia Naval y su organización se sujetará a lo establecido en la Tabla de Organización y Equipo TOE (No. 3-03-06-07-00-21).

El COCIB se fundamenta en dar respuesta a la necesidad de fortalecer y desarrollar operaciones para la defensa y seguridad nacional particularmente en el ciberespacio. Centrando su atención en aquellas

[2]Armada Nacional de Colombia. (2021). Armada de Colombia. Recuperado el 20 de Mayo de 24 de https://www.armada.mil.co/node/64757

2. [3]Ministerio de Defensa Nacional. (2021). Resolución No. 127 de 2021. Recuperado el día 20 de mayo. de 24 de http://marinanet.armada.mil.co

que salvaguarden la seguridad cibernética de la Armada Nacional, protegiendo sus activos digitales y garantizando la operación continua. Sus funciones principales se describen a continuación.

- **Monitoreo y Análisis:** Supervisa de manera continua los sistemas de información y comunicaciones de la Armada para detectar y analizar posibles amenazas cibernéticas.
- **Respuesta a Incidentes:** Coordina la respuesta ante incidentes cibernéticos, proporcionando asistencia técnica y operativa para mitigar los impactos y recuperar la normalidad en caso de ataques o intrusiones.
- **Protección de la Infraestructura:** Implementa medidas de seguridad cibernética para proteger la infraestructura crítica de la Armada, incluyendo sistemas de Comando y Control, comunicaciones, y otros activos digitales.
- **Gestión de Riesgos:** Evalúa y gestiona los riesgos cibernéticos, identificando vulnerabilidades y estableciendo medidas preventivas para garantizar la integridad, confidencialidad y disponibilidad de la información

Debido a la necesidad de una comprensión más profunda, el equipo COCIB y los autores del proyecto han establecido una colaboración bilateral para enfocar y desarrollar su entendimiento.

Analítica de Datos e Inteligencia Artificial en Ciberseguridad

Ciencia de Datos

La ciencia de datos es un campo interdisciplinario que utiliza métodos, procesos, algoritmos y sistemas científicos para extraer conocimiento de datos estructurados y no estructurados. Se basa en principios y técnicas de diversas disciplinas, como matemáticas, estadística, informática y disciplinas específicas del dominio. La ciencia de datos es fundamental para abordar problemas complejos en diversas industrias, incluyendo la ciberseguridad, donde se utiliza para detectar y mitigar amenazas.

Componentes de la Ciencia de Datos

La ciencia de datos se compone de varios componentes clave:

- **Recolección de Datos:** Es el proceso de reunir información de diversas fuentes. En el contexto de ciberseguridad, los datos pueden provenir de registros de eventos de seguridad, sistemas de detección de intrusos, firewalls, y otros dispositivos de red.
- **Preprocesamiento de Datos:** Involucra la limpieza y transformación de datos para asegurar su calidad y facilitar el análisis. Incluye la eliminación de datos duplicados, el manejo de valores faltantes y la normalización de datos.

- **Análisis Exploratorio de Datos (EDA):** EDA es una fase crítica en la que se exploran los datos para descubrir patrones, detectar anomalías, probar hipótesis y verificar supuestos. Las visualizaciones de datos y las técnicas estadísticas son herramientas comunes en esta fase.
- **Modelado Predictivo:** Utiliza algoritmos de aprendizaje automático (Machine Learning) para construir modelos que pueden predecir resultados futuros basados en datos históricos. Los modelos predictivos en ciberseguridad pueden identificar comportamientos anómalos y prever posibles amenazas.
- **Evaluación de Modelos:** Involucra la validación de la precisión y efectividad de los modelos predictivos mediante el uso de métricas como precisión, recall, y F1-score. La validación cruzada es una técnica comúnmente utilizada para evaluar modelos.
- **Implementación y Monitoreo:** La implementación de modelos en entornos operativos y su monitoreo continuo son esenciales para asegurar que los modelos sigan siendo efectivos ante nuevas amenazas. Esto también implica la actualización y reentrenamiento de los modelos basados en datos nuevos.

Algoritmos de Aprendizaje Automático

En el contexto de la ciberseguridad, varios algoritmos de aprendizaje automático pueden ser aplicados para la detección y mitigación de amenazas:

- **Árboles de Decisión:** Son modelos predictivos que utilizan un conjunto de reglas para hacer predicciones basadas en las características de los datos. Son fáciles de interpretar y pueden manejar datos categóricos y numéricos.
- **Máquinas de Vectores de Soporte (SVM):** Son modelos supervisados que se utilizan para clasificación y regresión. Son efectivos en espacios de alta dimensionalidad y se utilizan para detectar comportamientos anómalos en la ciberseguridad.
- **Redes Neuronales:** Son modelos inspirados en la estructura del cerebro humano. Las redes neuronales profundas (deep learning) son especialmente útiles para detectar patrones complejos y sutiles en grandes volúmenes de datos.
- **Clustering:** Es una técnica de aprendizaje no supervisado que agrupa datos en clusters basados en la similitud de sus características. El clustering puede ser utilizado para identificar comportamientos anómalos que no se alinean con patrones normales.

Aplicaciones de la Ciencia de Datos en Ciberseguridad

La ciencia de datos se aplica en ciberseguridad de las siguientes maneras:

- **Detección de Intrusos:** Utiliza modelos predictivos para identificar accesos no autorizados y actividades sospechosas en tiempo real.
- **Análisis de Malware:** Aplica técnicas de aprendizaje automático para clasificar y detectar malware basándose en características de comportamiento y firmas.
- **Monitoreo de Redes:** Utiliza análisis de datos para supervisar el tráfico de red, identificar anomalías y prevenir ataques DDoS.
- **Análisis de Riesgos:** Evalúa la vulnerabilidad de sistemas y redes, y predice posibles vectores de ataque para priorizar medidas de seguridad.

Desafíos y Futuro de la Ciencia de Datos en Ciberseguridad

La Ciencia de Datos enfrenta varios desafíos en el campo de la ciberseguridad, incluyendo la necesidad de grandes volúmenes de datos etiquetados, la adaptación a amenazas en constante evolución y la integración con sistemas de seguridad existentes. Sin embargo, el futuro de la ciencia de datos en ciberseguridad es prometedor, con el desarrollo continuo de algoritmos más avanzados y la mejora de las capacidades de procesamiento de datos.

En resumen, la Ciencia de Datos ofrece herramientas poderosas para mejorar la detección y mitigación de amenazas cibernéticas, contribuyendo significativamente a la seguridad de las organizaciones en el entorno digital actual.

Inteligencia artificial

La Inteligencia Artificial (IA) es un campo de la informática que se enfoca en crear sistemas capaces de realizar tareas que normalmente requerirían inteligencia humana, como el aprendizaje, la resolución de problemas y la toma de decisiones. La IA puede analizar grandes cantidades de datos, identificar patrones y anomalías, y aprender de experiencias pasadas para mejorar su rendimiento

En este sentido este busca aprovechar las implementaciones de la IA para mejorar significativamente las capacidades de ciberseguridad en el comando. La IA puede ayudar a analizar grandes cantidades de datos recopilados de diversas fuentes, como registros de eventos, tráfico de red y comportamiento de los usuarios. Esto permite identificar patrones y tendencias que pueden indicar amenazas potenciales. Además, la IA puede ayudar a desarrollar modelos predictivos que pueden predecir la probabilidad de futuros ataques, lo que permite al Comando tomar medidas proactivas para proteger sus sistemas y datos

Estado del arte

La ciberseguridad y la ciberdefensa se han convertido en aspectos cruciales para la seguridad nacional en muchos países, y Colombia no es una excepción. El documento CONPES 3701 de Colombia, reconociendo la creciente sofisticación de los ataques cibernéticos y la necesidad de proteger al Estado y su infraestructura crítica, busca establecer una política nacional integral de ciberseguridad y ciberdefensa para contrarrestar el aumento de amenazas informáticas. Este documento propone un enfoque multidimensional que abarca la creación de nuevas instituciones, el fortalecimiento de la cooperación interinstitucional, la capacitación especializada y la mejora de la legislación y la cooperación internacional en materia de ciberseguridad.

En este sentido, se hace un diagnóstico de la situación actual, identificando debilidades en la capacidad del Estado para enfrentar estas amenazas, la falta de coordinación interinstitucional y la necesidad de fortalecer la legislación y la cooperación internacional en esta materia.

Para abordar estos problemas, se propone la creación de varias instancias, incluyendo una Comisión Intersectorial para establecer políticas, un Grupo de Respuesta a Emergencias Cibernéticas (colCERT) para coordinar acciones, un Comando Conjunto Cibernético (CCOC) para la defensa y un Centro Cibernético Policial (CCP) para la seguridad.

Además, se hace hincapié en la importancia de la capacitación especializada en ciberseguridad y ciberdefensa, así como en el fortalecimiento de la legislación y la cooperación internacional en estas áreas. El documento también incluye un plan de acción detallado y un presupuesto para la implementación de estas políticas

El artículo "Ciberseguridad, un desafío para las Fuerzas Militares colombianas en la era digital" (Peña Suárez, 2023) analiza los retos que enfrentan las Fuerzas Militares de Colombia en el ámbito de la ciberseguridad. El autor destaca la importancia de este tema debido a la creciente dependencia de las tecnologías digitales y al aumento de las amenazas en el ciberespacio, haciendo especial enfasis es fundamental que las Fuerzas Militares colombianas desarrollen estrategias y líneas de acción para aumentar sus capacidades defensivas y ofensivas en el ciberespacio, y que establezcan alianzas estratégicas con otros actores para proteger la infraestructura crítica del país.

Por otro lado, autores nacionales como Cujabante Villamil, Bahamón Jara, Prieto Venegas y Quiroga Aguilar (2020) profundizan en el desarrollo institucional de la ciberseguridad y la ciberdefensa en Colombia, haciendo énfasis en cómo este desarrollo ha impactado las relaciones cívico-militares. Su análisis abarca la evolución de la política de seguridad digital en el país y la creciente participación de diversos actores, tanto civiles como militares, en la gestión del riesgo cibernético. Destacan

especialmente la redefinición de las relaciones entre civiles y militares en el ámbito de la ciberdefensa y la ciberseguridad.

Otro autor que aborda este tema es Cortés Borrero (2015), quien realiza un análisis detallado del estado actual de la política pública de ciberseguridad y ciberdefensa en Colombia. En su artículo, examina los desafíos emergentes en la era digital y la implementación de políticas públicas para hacer frente a estos desafíos, involucrando a diversos sectores de la sociedad.

Además, Urcuqui López, Navarro Cadavid, Osorio Quintero y García Peña (2018) exploran la aplicación de la ciencia de datos en el campo de la ciberseguridad en su libro "Ciberseguridad: un enfoque desde la ciencia de datos". Estos autores investigan la viabilidad de utilizar la ciencia de datos para desarrollar soluciones a problemas en ciberseguridad, utilizando proyectos de investigación específicos centrados en la detección de malware en dispositivos móviles y el control de defacement en páginas web como ejemplos ilustrativos.

Por otro lado, en un artículo de reflexión, Cortés Borrero (2015) analiza el estado actual de la política pública de ciberseguridad y ciberdefensa en Colombia. El autor examina los desafíos emergentes en la era digital y la implementación de políticas públicas para hacer frente a estos desafíos, involucrando a diversos sectores de la sociedad.

Por su parte E. Mayer et all (2014) discute los desafíos de analizar grandes volúmenes de datos de registro en entornos empresariales, donde la inconsistencia y la complejidad de los registros dificultan la detección de amenazas y menciona a Beehive, el cual utiliza un enfoque de aprendizaje no supervisado para identificar incidentes de comportamiento anómalo del host y agrupar hosts con comportamientos similares.

Por otra parte, Alazab, M., Khraisat, A., & Kumar, R. (2020) en Machine Learning-Based Cyber Threat Detection: A Comprehensive Review. Este artículo proporciona una revisión exhaustiva de los enfoques de detección de amenazas cibernéticas basados en aprendizaje automático, evaluando su eficacia, desafíos y tendencias emergentes.

Choudhary, A., Saini, J. K., & Singh, M. (2020) realizan un estudio exhaustivo que revisa varios algoritmos de aprendizaje automático utilizados para la detección de amenazas cibernéticas, destacando su efectividad en la identificación y mitigación de riesgos. En el mismo sentido Swami (2020) presenta un análisis detallado de las técnicas de aprendizaje automático aplicadas en el campo de la ciberseguridad, identificando sus aplicaciones prácticas y desafíos inherentes.

Finalmente, Ahmed, M., Mahmood, A. N., & Hu, J. (2021) presentan una investigación que examina el uso de técnicas de aprendizaje profundo en la detección de amenazas cibernéticas, destacando sus ventajas, limitaciones y áreas de aplicación.

Tecnologías y Herramientas Utilizadas

Rapid7 y R7 InsightDR

La herramienta empleada para la toma de muestras y generación de registros (Logs) es el Rapid7.

Rapid7 es una empresa especializada en la detección de amenazas cibernéticas. Sus soluciones analíticas recopilan, contextualizan y analizan datos de seguridad para ayudar a las organizaciones a mejorar su actitud frente al manejo de la ciberseguridad. Rapid7 proporciona información sobre el estado de seguridad de los activos y de los usuarios a través de redes virtuales, en la nube, móviles, privadas y públicas[4]. La herramienta Rapid7 ha permitido al comando recopilar datos relevantes desde el inicio hasta la conclusión de diversas actividades, brindando una perspectiva completa de las operaciones y transacciones llevadas a cabo en el entorno de la red de cómputo de esta institución. Estos registros incluyen información sobre eventos críticos, transacciones clave y patrones de comportamiento, proporcionando una visión detallada y estructurada de las operaciones diarias.

Rapid7 ofrece el R7 InsightDR, una avanzada solución de seguridad basada en la nube que va más allá de la simple detección de amenazas. Esta plataforma no solo identifica incidentes, sino que también responde de manera ágil y efectiva a ellos. Además, monitoriza la autenticación y brinda una visión completa de los endpoints, permitiendo a las organizaciones mantener un control total sobre su infraestructura de seguridad.

Una de las características destacadas de R7 InsightDR es su capacidad para unificar la gestión de vulnerabilidades con el análisis del comportamiento del usuario. Esto significa que no solo se detectan y se corrigen las vulnerabilidades en tiempo real, sino que también se analiza cómo interactúan los usuarios con el sistema, identificando posibles amenazas internas y patrones de actividad sospechosa. Esta combinación de funciones proporciona una defensa integral contra una amplia gama de ataques cibernéticos, abarcando desde intrusiones externas hasta amenazas internas.

MITRE y categorización de ataques comunes

MITRE es una organización sin fines de lucro que opera centros de investigación y desarrollo financiados por el gobierno de los Estados Unidos, conocidos como Centros de Investigación y

[4] Rapid7, (2024) Recuperado el 20 de may. de 24 de https://www.rapid7.com/

Desarrollo Financiados Federalmente (FFRDC). Su misión es abordar problemas críticos de interés nacional mediante la aplicación de tecnologías avanzadas y ciencias emergentes. MITRE colabora estrechamente con diversas agencias gubernamentales para ofrecer soluciones innovadoras en áreas clave como ciberseguridad, salud, defensa y seguridad nacional.

Una de sus contribuciones más destacadas en el ámbito de la ciberseguridad es el desarrollo del marco MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge), una herramienta ampliamente adoptada que proporciona un conocimiento detallado de las tácticas, técnicas y procedimientos (TTPs) utilizados por actores de amenazas. Este marco permite a las organizaciones mejorar su postura de seguridad cibernética al identificar y mitigar efectivamente las amenazas avanzadas. En este sentido, el comando se enfoca en aquellos ataques que ya han sido categorizados por MITRE ATT&CK.

1. Phishing (T1566)

El phishing es una técnica común en la que los atacantes envían correos electrónicos fraudulentos para engañar a los usuarios y hacerles revelar información confidencial o instalar malware.

2. Exploitation of Public-Facing Application (T1190)

Los atacantes explotan vulnerabilidades en aplicaciones web que son accesibles públicamente para obtener acceso a la red interna.

3. Valid Accounts (T1078)

Los atacantes utilizan credenciales válidas para acceder a sistemas y redes. Esto puede incluir el uso de credenciales robadas, adivinadas o predeterminadas.

4. PowerShell (T1059.001)

El uso de PowerShell es común debido a su capacidad para ejecutar comandos y scripts de forma remota y sin dejar muchos rastros evidentes.

5. Credential Dumping (T1003)

Los atacantes extraen credenciales almacenadas en sistemas comprometidos para moverse lateralmente y escalar privilegios.

6. Scheduled Task/Job (T1053)

La creación de tareas programadas permite a los atacantes ejecutar comandos o scripts de manera persistente en sistemas comprometidos.

7. Remote File Copy (T1105)

Los atacantes transfieren archivos maliciosos a sistemas comprometidos a través de métodos como SCP, FTP, SMB, y otros.

8. Command and Control (C2)

Las técnicas de comando y control permiten a los atacantes mantener comunicación con sistemas comprometidos y ejecutar comandos. Algunas técnicas comunes de C2 incluyen:

- **Application Layer Protocol (T1071):** Utilización de protocolos de capa de aplicación como HTTP, HTTPS, DNS para establecer comunicación C2.
- **DNS (T1071.004):** El uso de DNS para comunicaciones C2 puede incluir técnicas como la exfiltración de datos a través de consultas DNS.
- **Web Service (T1102):** Los atacantes utilizan servicios web legítimos para C2**.**
- **Remote Access Software (T1219):** Uso de software de acceso remoto como TeamViewer, VNC, RDP para mantener el control de los sistemas.

En general, existen diversos tipos de ataques que, según MITRE, ponen en riesgo la seguridad informática y los activos de las organizaciones. En este análisis, el enfoque será en aquellos que afectan directamente a los mandos medios y altos, como los administradores, clasificados como ataques de comando y control.

Metodología

La metodología ASUM-DM (Agile Software Usage Measurement and Data Mining) se presenta como una solución robusta para abordar los desafíos de ciberseguridad en el proyecto COCIB. Esta metodología, derivada de CRISP-DM, incorpora principios ágiles y técnicas de minería de datos, ofreciendo un marco de trabajo iterativo e incremental para el desarrollo de proyectos de análisis de datos en el ámbito de la ciberseguridad.

Fases Clave de ASUM-DM

1. **Planificación:** En esta fase inicial, se definen los objetivos específicos del proyecto COCIB, se identifican los stakeholders y se establece un plan detallado que incluye la definición del alcance, los recursos necesarios y los plazos establecidos. Se realiza un análisis exhaustivo del

contexto empresarial y los requisitos de seguridad para garantizar la alineación del proyecto con las necesidades de la organización.

2. **Recopilación de Datos:** Esta etapa se centra en la adquisición y preparación de los datos relevantes para el análisis de ciberseguridad. Se recopilan datos de diversas fuentes, como registros de eventos de seguridad, logs de sistemas y dispositivos de red, y se implementan mecanismos de seguridad para proteger la integridad y confidencialidad de los datos recopilados.
3. **Análisis de Datos:** En esta fase crucial, se aplican técnicas de minería de datos y análisis estadístico para identificar patrones, anomalías y posibles amenazas en los datos recopilados. Se utilizan algoritmos de aprendizaje automático para desarrollar modelos predictivos que permitan detectar y prevenir futuros incidentes de seguridad.
4. **Implementación:** Una vez que los modelos de seguridad han sido validados y refinados, se procede a su implementación en el entorno operativo de COCIB. Se establecen mecanismos de monitoreo continuo para evaluar la efectividad de las medidas de seguridad implementadas y se realizan ajustes según sea necesario para garantizar la protección continua de los sistemas y datos.

Ventajas de ASUM-DM

La adopción de ASUM-DM en el proyecto ofrece una serie de beneficios significativos:

- **Agilidad:** La naturaleza iterativa de ASUM-DM permite una rápida adaptación a los cambios en el panorama de amenazas y una respuesta ágil a los nuevos requisitos de seguridad.
- **Medición y Evaluación:** ASUM-DM enfatiza la medición continua del uso del software y los datos, lo que permite evaluar la efectividad de las medidas de seguridad implementadas y realizar mejoras basadas en datos concretos.
- **Integración de Minería de Datos:** La incorporación de técnicas avanzadas de minería de datos y aprendizaje automático permite un análisis más profundo de los datos, lo que facilita la detección temprana de amenazas y la identificación de vulnerabilidades ocultas.

Ataques de comando y control

El primer tipo de ataque seleccionado para el proyecto, de entre todo el conjunto de ataques, fue el de Comando y Control (C&C o C2). La elección específica de un tipo de ataque permite filtrar los DataFrames, seleccionando únicamente aquellos que contribuyan efectivamente al análisis de cada de los tipos de ataques.

¿Qué es Comando y Control?

Un ataque de tipo Comando y Control, también conocido como C2 o C&C, es un tipo de ciberataque en el cual un atacante obtiene el control remoto de uno o más sistemas informáticos infectados. Estas técnicas de ciberataque permiten al atacante controlar los sistemas infectados de forma remota, lo que puede utilizarse para diversos fines, como el robo de datos, la instalación de malware o el lanzamiento de otros ataques cibernéticos.

A través de esta conexión, los atacantes pueden enviar instrucciones a los equipos infectados, hacerlos realizar tareas maliciosas, robar información, etc.

Los servidores de C&C son utilizados por los atacantes para coordinar y controlar la actividad de la red de equipos infectados, llamada "botnet".

El C&C es la infraestructura y las técnicas que los atacantes usan para comunicarse con dispositivos comprometidos (zombies) dentro de una red objetivo. Les permite enviar comandos, descargar malware adicional y extraer datos robados.

¿Cómo funciona?

- **Canales encubiertos:** Los atacantes usan canales de comunicación ocultos, a menudo disfrazados de tráfico legítimo (HTTP/HTTPS, DNS) y cifrados para evitar ser detectados.
- **Plataformas C&C:** Pueden ser personalizadas o estándar (Cobalt Strike, Covenant, etc.).
- **Botnets:** Redes de dispositivos zombies controladas por un C&C para realizar tareas maliciosas como ataques DDoS o minería de criptomonedas.
- **Beaconing:** Proceso por el cual los dispositivos infectados se comunican con el C&C para recibir instrucciones.

Gráfica 1: Ruta de una Ataque de C2

¿Qué pueden hacer los atacantes con C2?

- **Movimiento lateral:** Moverse a través de la red de la víctima para encontrar objetivos de alto valor.
- **Ataques en varias fases:** Lanzar ataques más complejos y dirigidos.
- **Exfiltración de datos:** Robar información sensible de la red objetivo.
- **Otros usos:** Ataques DDoS, minería de criptomonedas, etc.

Modelos C2:

- **Centralizado:** Los dispositivos infectados se comunican directamente con un servidor C&C central.
- **Entre pares (P2P):** Los dispositivos infectados se comunican entre sí sin un servidor central.
- **Controlado externamente con selección aleatoria de canales:** Se utilizan canales inusuales como redes sociales o incluso exploración aleatoria de Internet para comunicarse.

Detección y bloqueo del tráfico C2:

- **Monitorear y filtrar el tráfico saliente:** Utilizar firewalls, proxies y servicios de filtrado DNS.
- **Estar atento a las balizas:** Utilizar herramientas como RITA o analizar manualmente el tráfico.
- **Llevar registros y realizar controles:** Analizar registros en busca de patrones inusuales.
- **Comparar y contrastar datos de distintas fuentes:** Utilizar herramientas como Varonis Edge para obtener una visión completa.

Análisis exploratorio de datos y estudio de Logs

El análisis de datos exploratorio o simplemente EDA es un enfoque analítico que se utiliza para comprender y analizar los datos de manera inicial y profunda. Es un enfoque flexible, que además debe ser iterativo para dialogar con los datos, en todo el proceso del ciclo de vida del proyecto.

A continuación, se presentan cada una de las fases del EDA aplicadas a los datos recopilados

Recolección de Datos

La muestra entregada por el COCIB proviene de distintos sensores ubicados de manera estratégica y que son registrados y almacenados con el rapid7; esta muestra consiste en dos grupos de archivos de distintas fechas, con los respectivos registros (logs) de cada una de ellas.

La primera muestra es del 24 de marzo del 2024 al 30 de marzo del 2024. La muestra cuenta un total de 5 tablas discriminadas a continuación:

File Name	Size
Active Directory	831 Kb
Asset Authenticathor	831 KB
Audit Logs	480227 KB
DNS Query	745996 Kb
File Access Activity	3149166 KB

La otra muestra es del 07 de abril del 2024 al 13 de abril del 2024. Esta muestra tiene un total de 18 tablas discriminadas de la siguiente manera:

File Name	Size
Active Directory	1896 KB
Asset Authentication	3128173 KB
Audit Logs	322 KB
DNS Query	773055 KB
Endpoint Activity	8013366 KB
File Access Activity	2709269 KB
File Modification Activity	336999 KB
Firewall Activity	2716502 KB
Host To IP Observations	1310100 KB
IDS Alert	325997 KB
Ingress Authentication	4019 KB
Internal Logs	7 KB
Network Flow	1249659 KB

Raw Logs	1570448 KB
Third Party Alert	54 KB
Unparsed Data	76010 KB
Virus Alert	102 KB
Web Proxy Activity	3126848 KB

Preprocesamiento de Datos

Una de las primeras etapas del preprocesamiento se llevó a cabo cuando se eligen, de todas las tablas entregadas por el comando, aquellas que aportaran de manera directa o indirecta al ataque elegido C2 o C&C

DataFrames para C&C

En este sentido, con el fin de llevar a cabo un análisis exhaustivo de los ataques que se ajusten al perfil de Comando y Control, se dará prioridad al estudio de las siguientes tablas y sus relaciones y correlaciones:

1. Active Directory: es una base de datos y un conjunto de servicios que conectan a los usuarios con los recursos de red que necesitan para realizar su trabajo. La base de datos (o directorio activo) contiene información crítica sobre su entorno, incluidos los usuarios y las computadoras que hay y quién puede hacer qué.
2. Asset Authenticator: El Asset Authenticathor es una herramienta que forma parte de la suite de productos de Rapid7 y se utiliza para realizar autenticación y autorización de activos (dispositivos, sistemas, aplicaciones, etc.) en un entorno empresarial.

   Algunas de las principales funcionalidades y características del Asset Authenticathor de Rapid7 son:

- Descubrimiento de activos: Permite identificar y catalogar todos los activos conectados a la red, incluyendo dispositivos, servidores, aplicaciones, etc.
- Autenticación de activos: Verifica la identidad de los activos a través de diversos métodos de autenticación, como credenciales, certificados digitales, etc.
- Gestión de accesos: Controla y administra los permisos y privilegios de acceso a los diferentes activos, asegurando que solo los usuarios y procesos autorizados puedan interactuar con ellos.

- Monitoreo y alertas: Realiza un seguimiento continuo de la actividad de los activos, generando alertas en caso de detección de comportamientos sospechosos o actividades no autorizadas.

3. DNS Query: En el contexto de Rapid7, el "DNS Query" es el conjunto de solicitudes de información enviada a un servidor DNS (Sistema de Nombres de Dominio) para traducir un nombre de dominio en una dirección IP. Rapid7 puede analizar estas consultas para detectar actividad sospechosa, como ataques cibernéticos o comunicación con servidores comprometidos
4. IDS Alert: El archivo IDS Alert de Rapid7 es un componente relacionado con la detección y respuesta a intrusiones (IDS, por sus siglas en inglés) en la infraestructura de TI de una organización. IDS, o Sistema de Detección de Intrusiones, es una herramienta de ciberseguridad que monitorea el tráfico de red o actividades del sistema en busca de comportamientos sospechosos o maliciosos.

   Algunas de las principales funcionalidades y características del IDS Alert de Rapid7 son:

   - Registrar Actividades Sospechosas: Almacena registros detallados de eventos o actividades que el sistema de detección considera potencialmente maliciosas o anómalas. Estos eventos pueden incluir intentos de acceso no autorizados, escaneos de red, intentos de explotación de vulnerabilidades, entre otros.

   - Análisis de Incidentes: Los datos almacenados en el archivo IDS Alert se utilizan para analizar incidentes de seguridad. Esto ayuda a los equipos de seguridad a investigar eventos sospechosos, determinar la naturaleza de las amenazas y tomar acciones correctivas.

   - Generación de Alertas: Basándose en los eventos registrados, el sistema puede generar alertas en tiempo real para notificar a los administradores de seguridad sobre posibles incidentes. Estas alertas son fundamentales para una respuesta rápida y eficaz.

   - Cumplimiento y Auditoría: Mantener un registro de todas las alertas de IDS es también importante para cumplir con diversas regulaciones de seguridad y auditorías. Proporciona una pista de auditoría que puede ser revisada en caso de un incidente de seguridad significativo.

5. File Access Activity: En el contexto de Rapid7, "File Access Activity" (Actividad de Acceso a Archivos) se refiere al monitoreo y registro de acciones relacionadas con archivos dentro de un entorno de red o sistema informático. Esta actividad incluye cualquier tipo de acceso que

implique acceder, leer, escribir, modificar, eliminar o mover a los diferentes tipos de archivo en una red o en sistemas locales, registrando datos como el usuario, dominio e IP desde el cual se realizó el acceso junto con los horarios en que se realizaron dichas actividades. Este log es crucial para la seguridad de la información y la detección de amenazas, ya que puede ayudar a identificar actividades sospechosas (atipicas) o maliciosas, como intentos de acceso no autorizado, cambios no autorizados en archivos críticos, intentos de exfiltración de datos o actividades en horarios irregulares.

6. File Modification Activity: En el contexto de Rapid7, "File Modification Activity" (Actividad de Modificación de Archivos) se refiere al monitoreo y registro de cambios realizados en archivos dentro de un entorno de red o sistema informático. Esta actividad incluye cualquier acción que implique la modificación, actualización, creación o eliminación de archivos en una red o en sistemas locales. La monitorización de la actividad de modificación de archivos es crucial para la seguridad de la información y la detección de amenazas, ya que puede ayudar a identificar actividades sospechosas o maliciosas, como cambios no autorizados en archivos críticos, manipulación de archivos por parte de usuarios no autorizados, o intentos de alterar o borrar datos importantes.
7. Firewall Activity: En el contexto de Rapid7, "Firewall Activity" (Actividad de Firewall) se refiere al monitoreo y registro de eventos y acciones relacionadas con el firewall de red. Esto incluye cualquier actividad que implique el tráfico de red que pasa a través del firewall, como conexiones entrantes y salientes, reglas de firewall aplicadas, intentos de conexión bloqueados o permitidos, y otros eventos relacionados con la seguridad de la red.

   El firewall es una parte fundamental de la infraestructura de seguridad de una red, ya que actúa como una barrera entre la red interna y externa, controlando y filtrando el tráfico de red según las reglas predefinidas. Monitorear la actividad del firewall es esencial para detectar y prevenir posibles amenazas de seguridad, como intentos de intrusión, tráfico malicioso, o comportamientos anómalos que podrían indicar un ataque.

Limpieza general de Datos

El proceso de limpieza de datos, crucial en todo el proceso, implica identificar y corregir errores, valores faltantes, duplicados y otros problemas que puedan eventualmente afectar la calidad de los resultados.

En este sentido las tablas seleccionadas para C2 se sometieron al Datapipeline (anexo). Esta limpieza se llevó a cabo para todas las tablas e incluye aplicar dos criterios generales:

Criterio 1: En todas las tablas, se eliminarán aquellas columnas cuyos registros presenta un 70% o más de valores faltantes o nulos. El método drop () aplicado en pipeline permite automatizar el proceso en todas las tablas.

Criterio 2: En algunas columnas de las tablas se presentan un alto número de datos duplicados que genera desbalanceo, en algunos casos incluso el 100%, lo cual afecta el análisis de datos. En este sentido aquellos cuyos registros tengan el 60% de datos iguales se eliminan, tomando atenta y cuidadosa nota de su significado en las fases finales.

Una vez se lleva a cabo la limpieza, procedemos a hacer las transformaciones en atributos y registros.

Transformaciones y selección de atributos

Duplicidad de columnas: Se hizo muy evidente en la mayoría de las tablas que algunas de sus columnas se comportaban de manera muy similar, teniendo una especie de duplicidad entre ellas, si bien en algunos casos no era del 100% de registros, se elige el criterio que aquellas columnas que tengan el 80% o más de datos iguales, se eliminen dejando una, que actuará como representante de las demás. Este proceso busca además de evadir la correlación, reducir la dimensionalidad de las tablas. Para esto se tuvo en cuenta el estadístico Xhi cuadrado.

Datatime: En todas las tablas se hace la transformación del formato timestamp a datatime. Esto se debe a que todas las fechas de las tablas en el cual se hizo el registro del evento o logs están en formato ISO 8601. Estos datos tienen la Zona horaria zulú: La Z al final indica que las marcas de tiempo están en Tiempo Universal Coordinado (UTC). Para convertirla a la hora local de Colombia (GMT-5), se debe restar 5 horas y hacer la transformación a un formato de YYYY-MM-DD HH: MM.

Manejo de Datos Faltantes: Para el manejo de datos faltantes, se aplicarán de manera simultánea dos de los métodos más usados

- **Eliminación (listwise deletion):** En algunas tablas se optará por eliminar las filas, esto se debe a que, por la naturaleza de los datos, se hace difícil anticipar el dato faltante, por lo que no se puede aplicar algún tipo de imputación
- **Imputación**: usando métodos basados en estadísticas y modelado, se buscará imputar los datos en algunas de las tablas cuya naturaleza permita aplicar el método.

Una vez los Dataframe se someten a esta limpieza y transformación general, se lleva a cabo un proceso adicional de selección de variables (Feauture Selection) a cada una de las tablas (Ver anexo), entendiendo que cada tabla tiene sus características propias, este proceso se hizo combinando estadísticos de prueba, como xhi cuadrado con niveles de confiablidad del 0.95 y un nivel de

significancia de 0.05 además de visión de expertos en el negocio. Esto se hace con el objetivo de buscar reducir dimensionalidad,

Interpretación de los Resultados

El archivo **Active Directory** se origina a partir de la recopilación de registros de seguridad del controlador de dominio. Estos registros contienen información detallada sobre eventos importantes como inicios de sesión, cambio de política y administración de cuentas. Este archivo permite llevar un seguimiento detallado de inicios de sesión fallidos en cuentas específicas, facilitando la detección proactiva de las posibles actividades sospechosas dentro del entorno de red.

La base de datos original contenía 906 registros y 75 columnas. Tras el proceso de limpieza de datos, se detectó que 39 de estas columnas tenían más del 70% de valores nulos. Siguiendo las recomendaciones recibidas, decidimos eliminar estas columnas para mejorar la calidad del análisis. A pesar de esto, persiste un porcentaje variable de valores nulos entre el 5% y el 65% en algunas columnas restantes, como se muestra en la Ilustración 1. En lugar de imputar estos valores, se optó por rellenarlos con el término "desconocido" en algunas columnas.

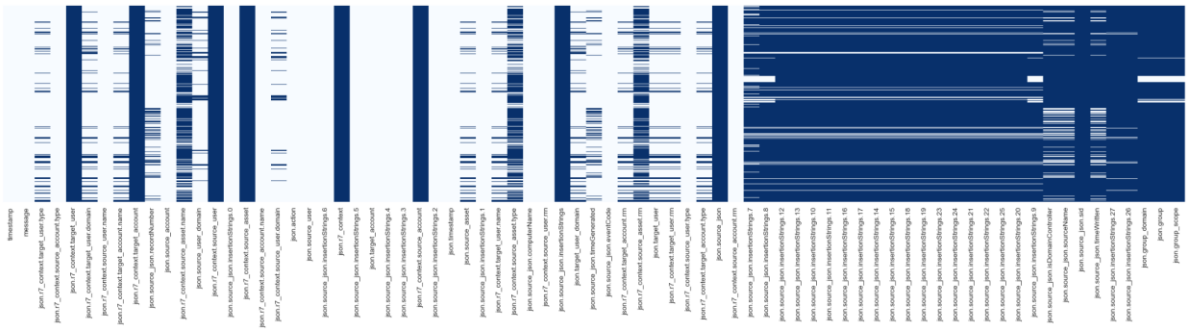


Ilustración 1: Mapa de calor de los valores nulos del archivo Active Directory

Además, identificamos 4 columnas con un desbalance significativo, donde una categoría representaba la mayoría de los registros. Estas columnas también fueron eliminadas para mantener la integridad del análisis. También encontramos casos de columnas con información duplicada, como en el caso de 'json.target_user_domain' y 'json.r7_context.target_user.domain', así como 'json.r7_context.target_user.name' , ‘json.source_json.insertionStrings.0’ y 'json.target_account', estas 3 columnas contenían la misma información del nombre de usuario, entre otras duplicidades. Finalmente, el archivo quedó con 28 columnas preparadas para un análisis detallado.

	column	value	percentage
0	json.r7_context.target_user.type	user	86.754967
0	json.r7_context.source_account.type	account	100.000000
0	json.r7_context.source_user.type	user	100.000000
0	json.r7_context.target_account.type	account	86.754967

Ilustración 2: Columnas con valores únicos del archivo Active Directory

Durante el análisis exploratorio, observamos que de los 906 registros, 195 pertenecían al dominio Carreño y 10 al dominio Tumaco. Esto proporcionó una visión clara de la distribución de los datos por dominio."desconocido".

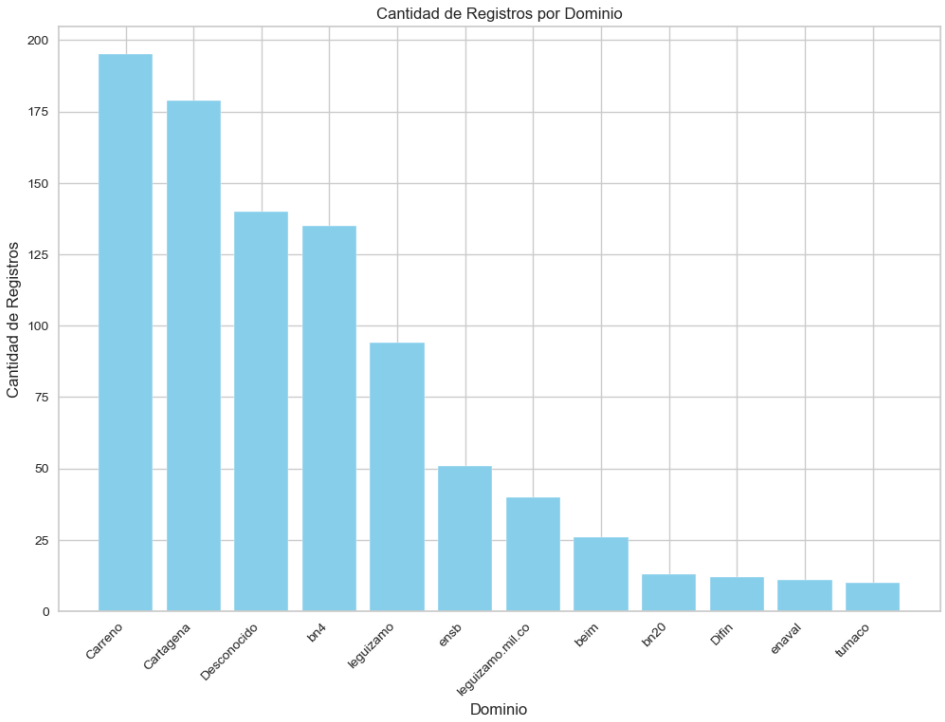


Ilustración 3: Cantidad de registros por dominio del archivo Active Directory

El análisis adicional de cada código del atributo "json.source_json.eventCode" se tradujo usando el diccionario de eventos de Rapid7. El evento más frecuentemente registrado fue el de "cuenta de usuario bloqueada", con 288 registros, seguido por "intento de restablecer contraseña por administrador", con 189 registros. En contraste, el evento menos común fue el "cambio de cuenta de usuario", que solo tuvo 2 registros.

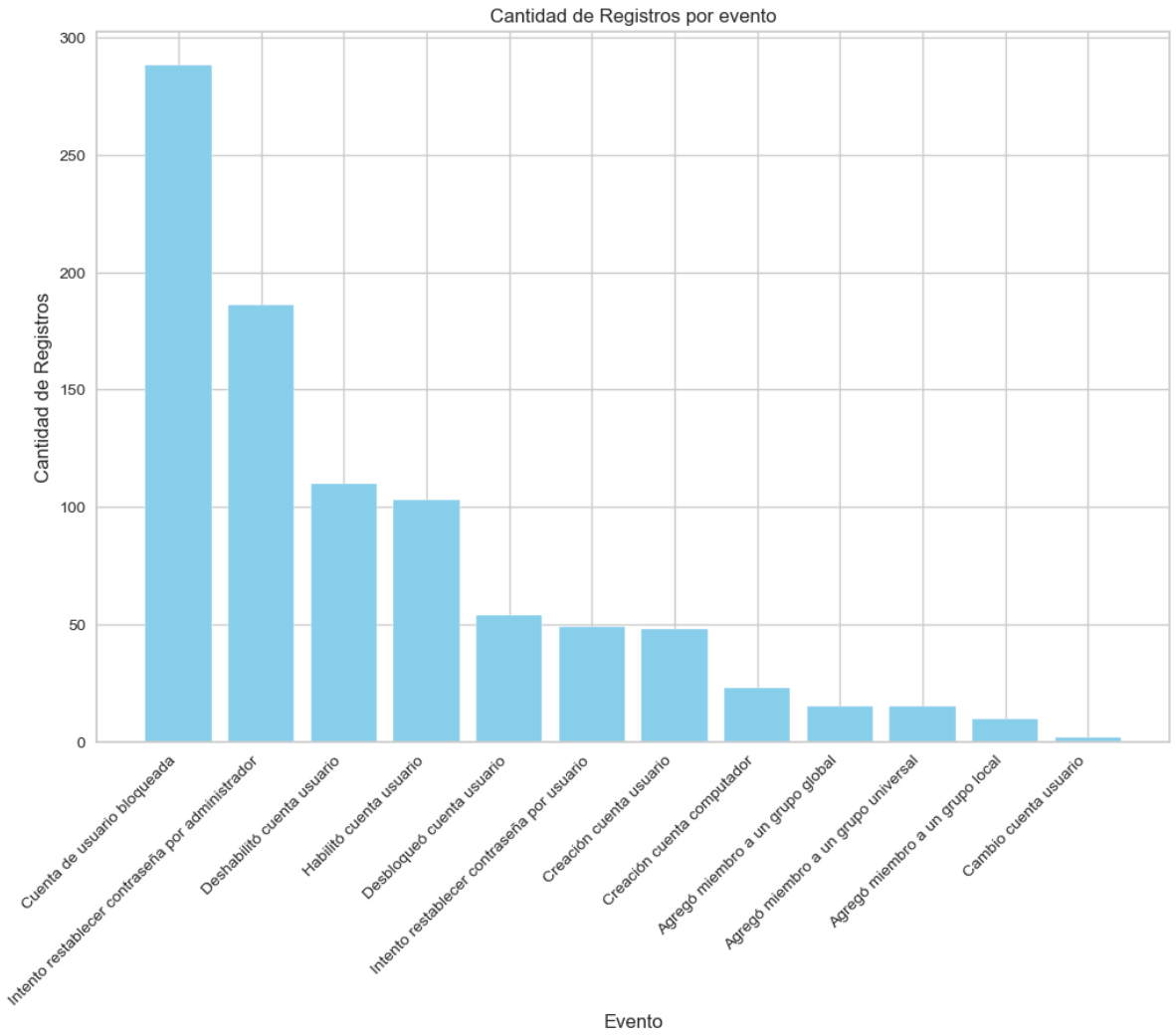


Ilustración 4: Cantidad de registros por eventos del archivo Active Directory

En el Asset Authentication como se ha mencionado, contiene toda la data relacionada con los accesos a los diferentes equipos de la armada, además del acceso a paginas externas, en este archivo se pueden ver particularmente dos registros muy importantes para la organización: los IP de los diferentes equipos y los puertos mas usados.

En este sentido, se hizo un análisis por cantidad de registros según las fechas

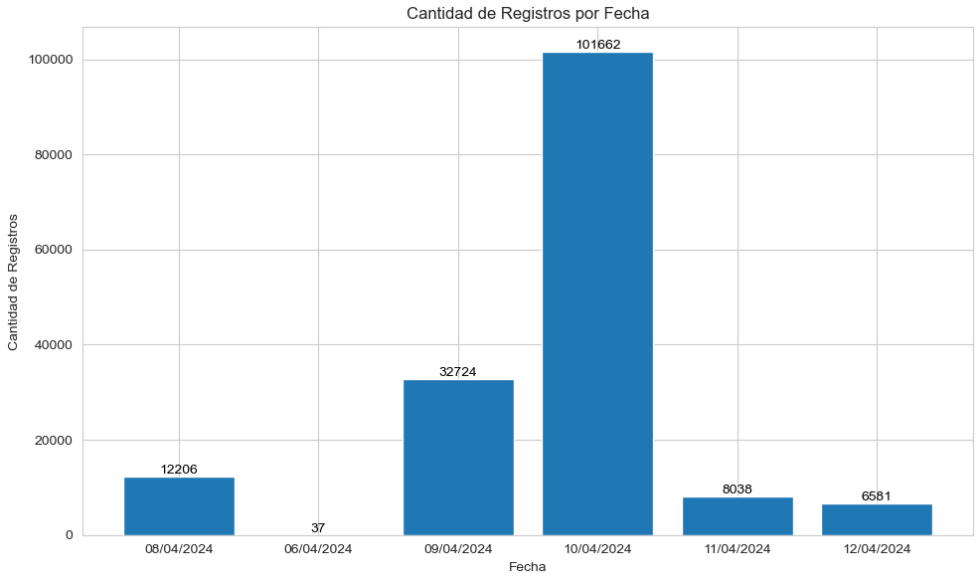


Ilustración 1: Cantidad de registros por fechas

Asi mismo se evidencia un registro de entradas por horas

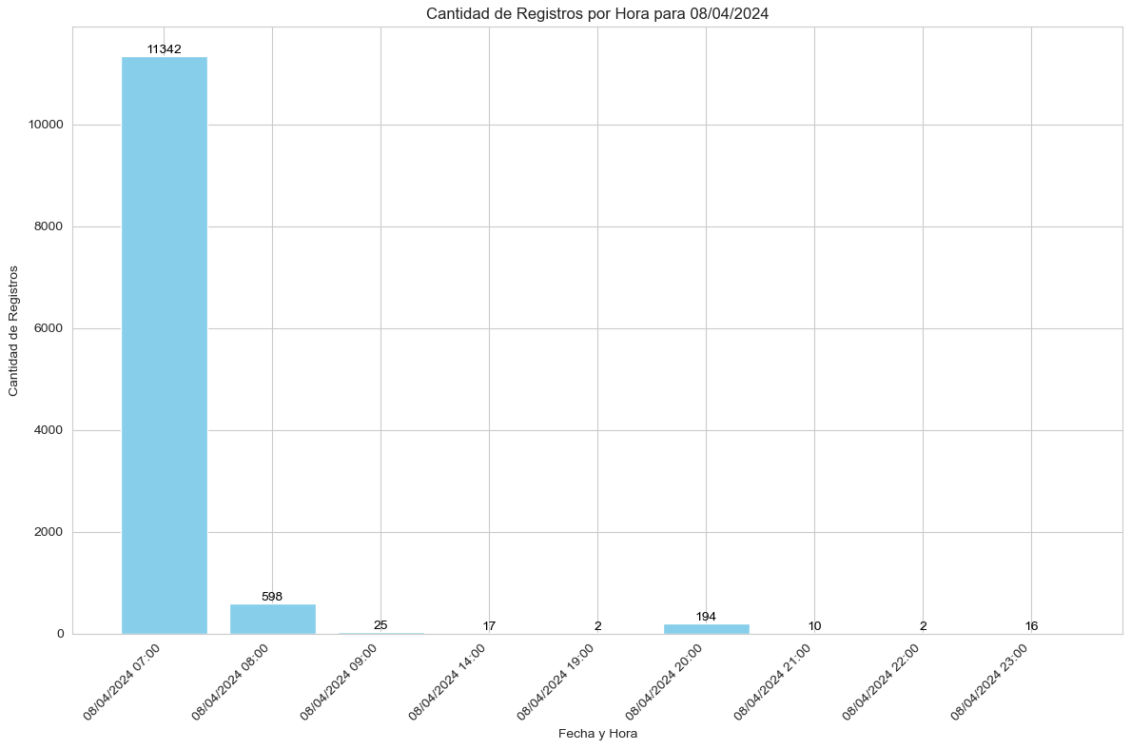


Ilustración 2: Ingresos por horas

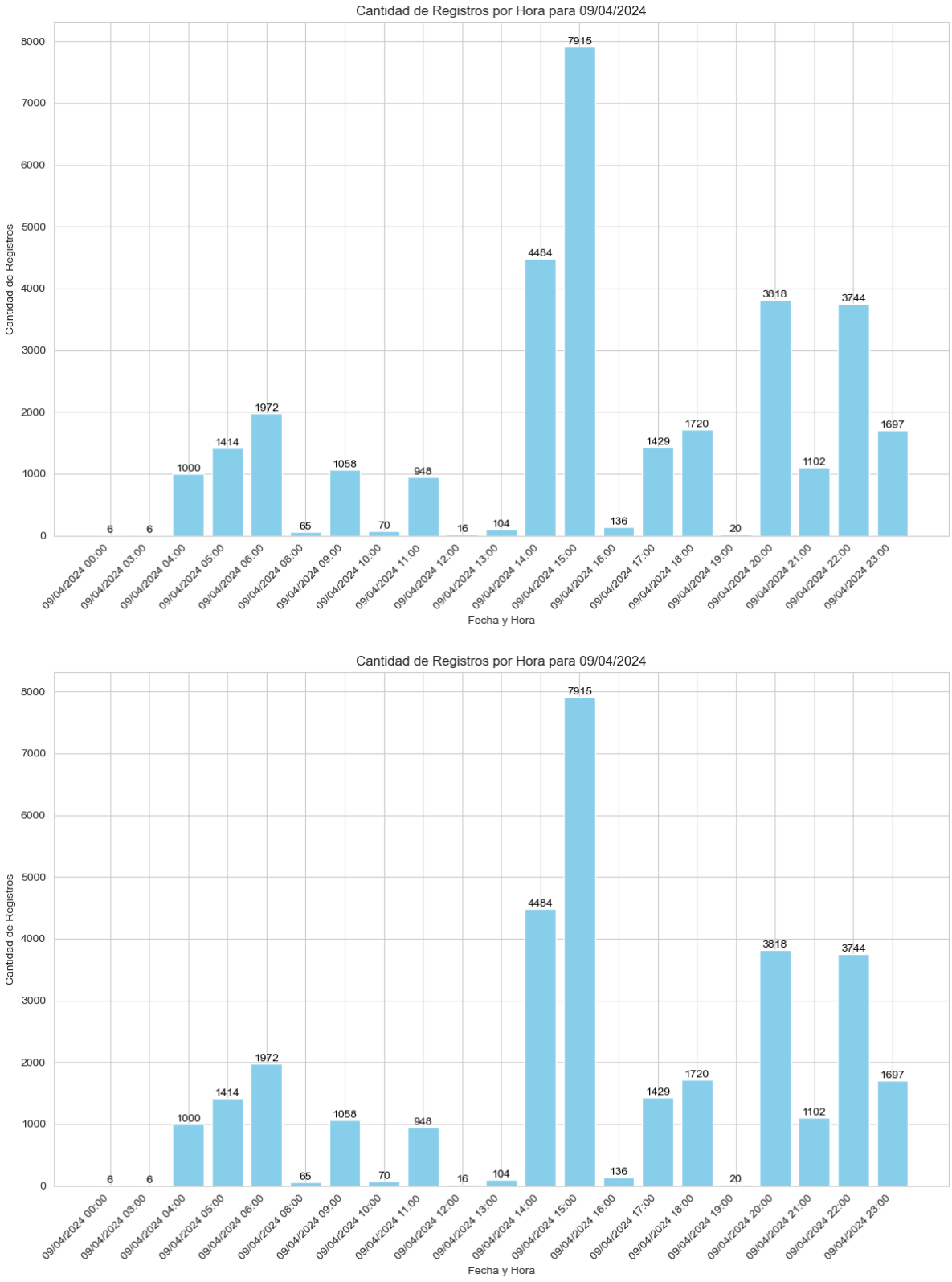
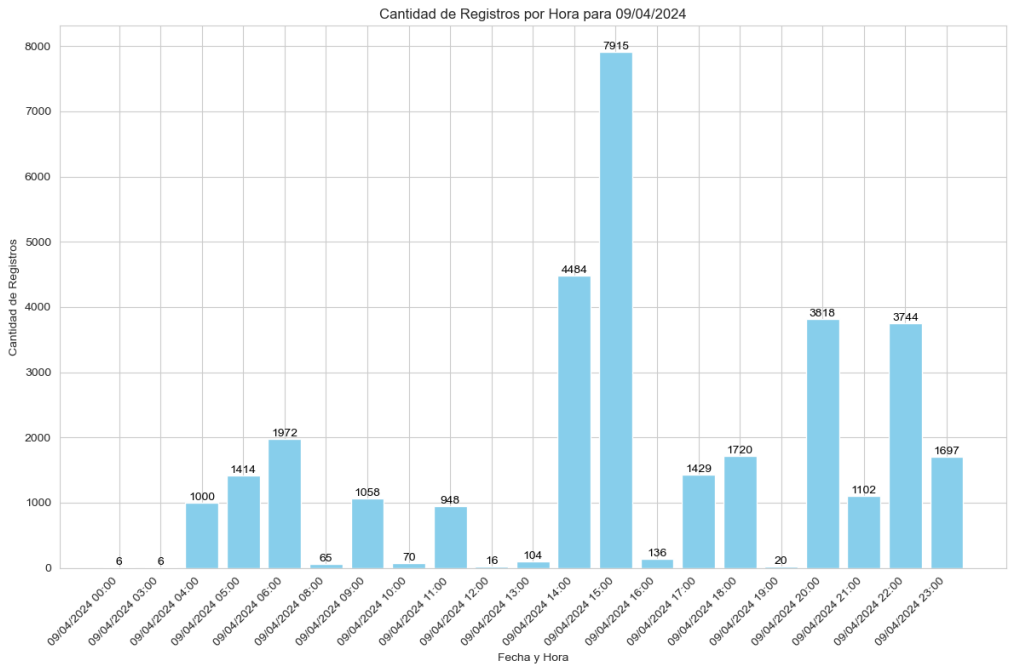


Ilustración 3: ingreso por hora

Analizar el ingreso por horas en algunas fechas nos permite analizar como es el comportamiento normal en un día, ya que nuestro objetivo principal en el Asset es identificar anormalidad, es indispensable verificar dichos comportamientos normales para poner la lupa en aquellos que sean atípicos en un día del comando.



En algunos de estos gráficos nos encontramos con el comportamiento típico de los ingresos y registros por horas. Su comportamiento varía de acuerdo con el día, pero los registros se toman las 24 horas. Hay momentos en el día que no hay muchas variaciones, pero en el horario habitual que podría desde las 8:00 hasta las 16:00 los ingresos, aunque variados suelen ser habitual.

Uno del aspectos mas relevantes en el Asset fue la búsqueda de la correlación entre columnas, como ya menciono anteriormente la eliminación de columnas por el criterio de nulos y luego por el criterio de semejanzas nos brindó un nuevo archivo con alrededor de 10 nuevos atributos, en este sentido la grafica de correlación entre estos nuevos atributos es:

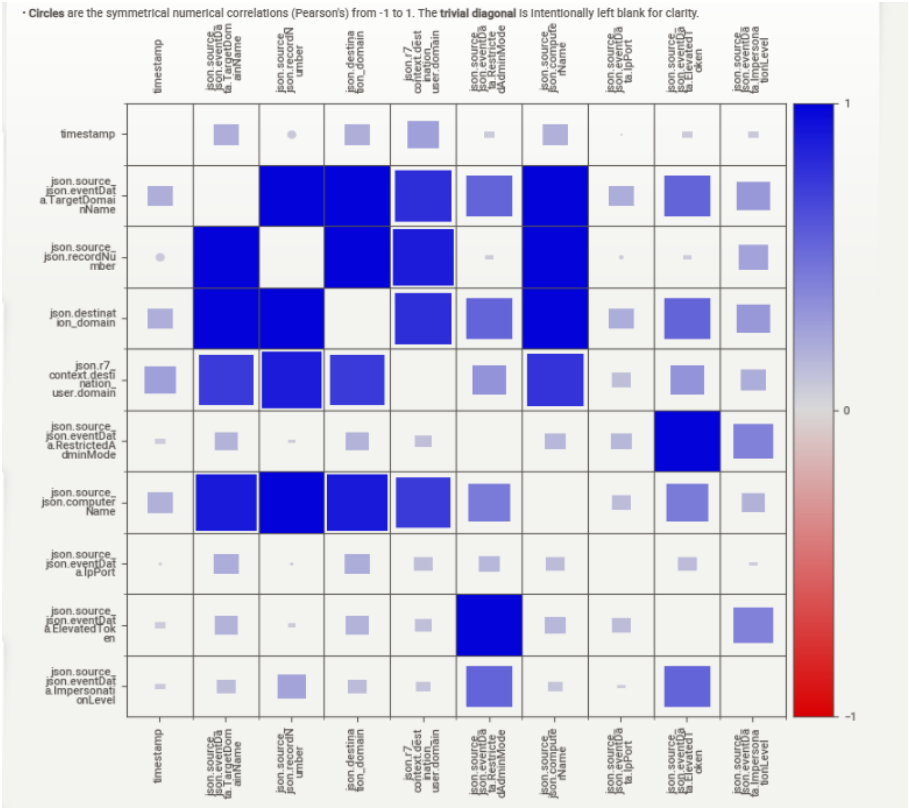


Ilustración 4: Mapa de Correlación entre atributos

Estas nuevas correlaciones permiten verificar efectos de causalidad y casualidad en los registros y tomar decisiones en pro de la ejecución del modelo

Otro aspecto importante que será analizado a futuro es el dominio desde donde y hacia donde se hacen los registros de actividades en las fechas. Mayoritariamente se hacen en Bogotá y Cartagena.

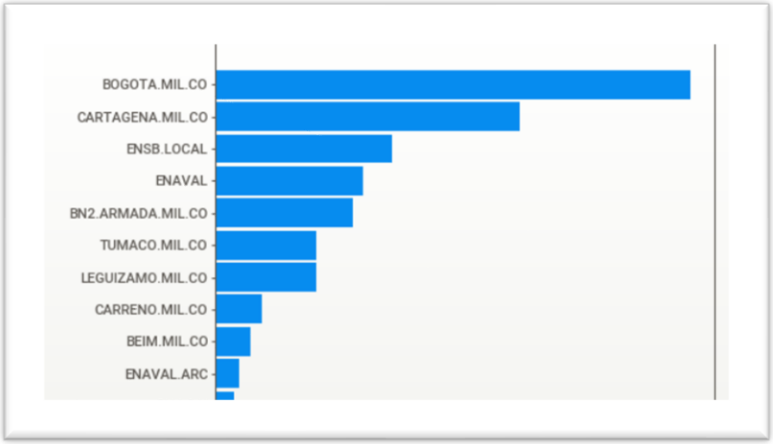


Ilustración 5: Dominio del Asset

El archivo DNS QUERY como se mencionó anteriormente, hace referencia al conjunto de solicitudes de información enviadas a un servidor DNS (Sistema de Nombres de Dominio) para traducir un nombre de dominio en una dirección IP. Cada vez que se realiza esta traducción, se genera un log o registro de esta actividad. Estos logs de DNS son una herramienta vital para la seguridad, el monitoreo y la gestión efectiva de redes. Ayudan a detectar y prevenir amenazas, solucionar problemas y asegurar que la red funcione de manera eficiente y segura.

Entendamos mejor esto con un ejemplo. Imaginemos que empezamos a notar un aumento repentino en el tráfico de red. Al revisar los logs de DNS, descubrimos que muchos dispositivos están consultando un dominio extraño que nunca se había visto. Investigando más, encontramos que este dominio está relacionado con un nuevo tipo de malware. Gracias a los logs de DNS, podemos actuar rápidamente para bloquear ese dominio y prevenir una infección mayor en la red.

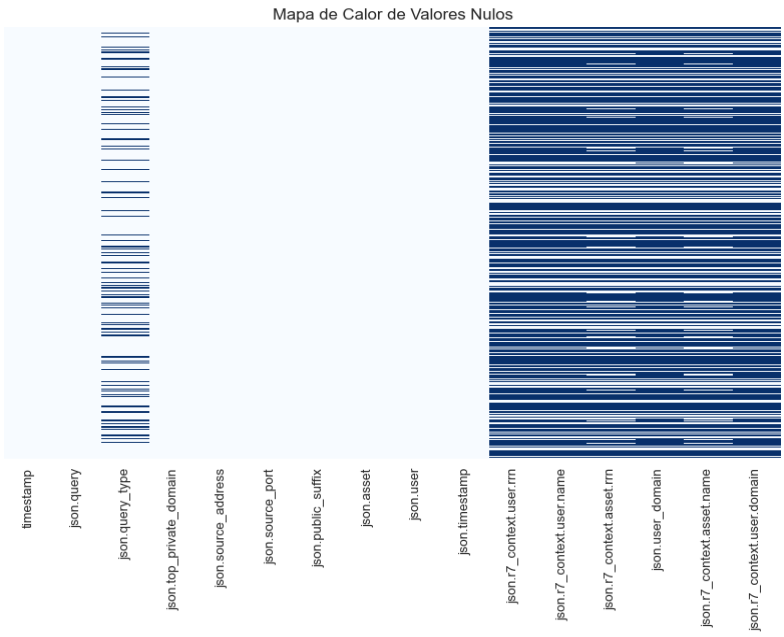


Ilustración 1: Mapa de calor de los valores nulos del archivo DNS Query

Siguiendo ahora con la información proporcionada, la base de datos contenía un millón de registros y un total de 28 columnas. Una vez que el archivo se pasó por el pipeline, se detectó que 6 de las columnas tenían más del 70% de valores faltantes. Por lo tanto, siguiendo la recomendación de nuestros mentores, decidimos eliminar estas columnas.

Aun así, la base de datos quedó con un porcentaje de valores nulos que oscilaba entre el 5% y el 65% (ver Ilustración 1). No eliminamos estas columnas restantes; en cambio, dado que no consideramos viable imputar los valores, decidimos rellenarlos con la palabra "desconocido".

Adicionalmente, había 5 columnas que presentaban desbalanceo, es decir, una mayor proporción de una sola categoría. Estas columnas fueron eliminadas. Al final, el archivo quedó con 16 columnas listas para el análisis.

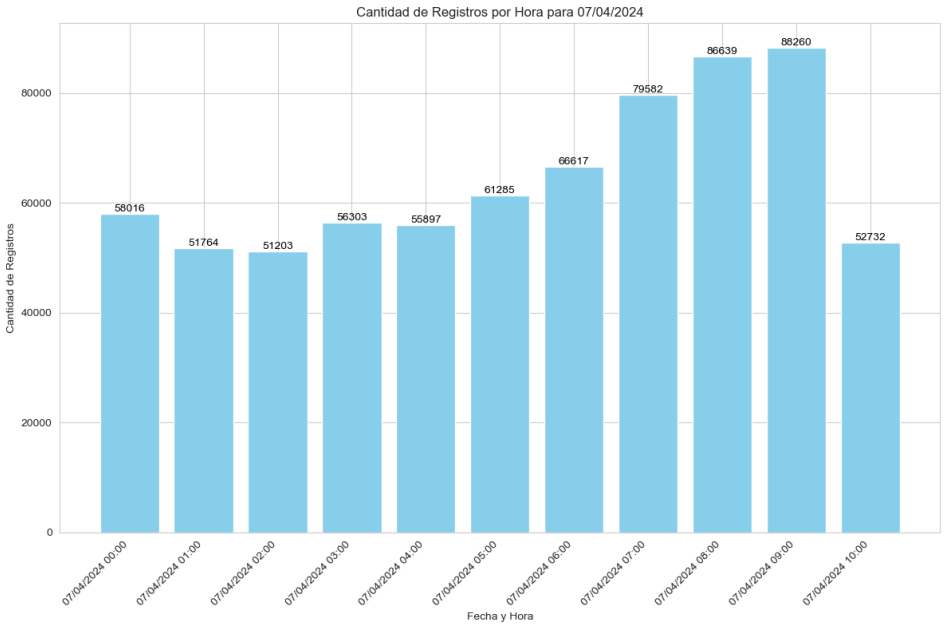


Ilustración 2: Cantidad de registros por hora para el 07 de abril de 2024

La base de datos solo tenía información de dos días y no contenían todas las horas del día. Por ejemplo, como se muestra en la Ilustración 2, el día con más información solo tenía registros entre las 12 de la noche y las 10 de la mañana. En este caso, la mayor cantidad de registros o logs se presentaba entre las 7 y 9 de la mañana, que son las horas en las que los empleados comienzan su jornada laboral y, por lo tanto, consultan diversos dominios en la red.

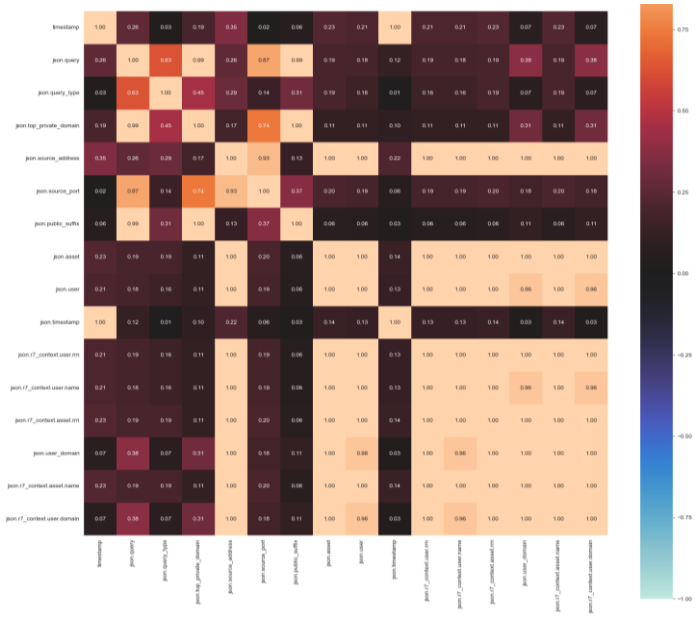


Ilustración 3: Gráfico de correlaciones entre las variables del archivo DNS Query

Adicionalmente, en la Ilustración 3 se puede observar el gráfico de correlaciones, en el cual se nota que hay una fuerte relación entre varias de las columnas. Es importante evaluar si debemos eliminar estas variables, ya que tener variables altamente correlacionadas puede traer problemas como la multicolinealidad al momento de elaborar un modelo. Además, mantener ambas variables puede resultar redundante.

En el archivo **File Access Activity**, se registran las operaciones de acceso sobre los diferentes archivos, en los diferentes objetos de infraestructura crítica dentro de la red monitoreada con el aplicativo **Rapid7**; Como se mencionó previamente, el archivo File Access Activity guarda información del equipo, el usuario al igual que su dominio, la ruta , extension del archivo y tipo de acceso que se ejecuta sobre el mismo.

Columna	Nulos	Porcentaje Nulos	Valores Distintos		Total Registros	1000000
timestamp	0	0,000%	15888			
json.file_path	0	0,000%	119669			
json.source_json.isDomainController	0	0,000%	2			
json.source_json.insertionStrings.0	0	0,000%	55			
json.file_share	0	0,000%	10			
json.source_address	0	0,000%	32			
json.source_json.insertionStrings.8	0	0,000%	10			
json.source_json.insertionStrings.7	0	0,000%	10			
json.source_json.insertionStrings.6	0	0,000%	2928			
json.source_json.insertionStrings.5	0	0,000%	28			
json.source_json.insertionStrings.4	0	0,000%	1			
json.source_json.insertionStrings.3	0	0,000%	3919			
json.source_json.insertionStrings.2	0	0,000%	2			
json.timestamp	0	0,000%	967487			
json.source_asset	0	0,000%	34			
json.source_json.insertionStrings.1	0	0,000%	57			
json.target_address	0	0,000%	7			
json.source_json.computerName	0	0,000%	7			
json.file_name	0	0,000%	43104			
json.source_json.insertionStrings	1000000	100,000%	0			
json.source_json.insertionStrings.9	0	0,000%	119655			
json.source_json.sourceName	0	0,000%	1			
json.access_types	0	0,000%	23			
json.source_json.sid	1000000	100,000%	0			
json.source_json.eventCode	0	0,000%	1			
json.source_json.insertionStrings.12	0	0,000%	57			
json.source_json.insertionStrings.10	0	0,000%	34			
json.source_json.insertionStrings.11	0	0,000%	34			
json.service	0	0,000%	1			
json.source_json.timeWritten	0	0,000%	999858			
json.file_extension	0	0,000%	2433			
json.source_json	1000000	100,000%	0			
json.user	0	0,000%	31			
json.account	0	0,000%	57			
json.r7_context.source_asset.name	3445	0,345%	33			
json.r7_context.source_asset	1000000	100,000%	0			
json.r7_context	1000000	100,000%	0			
json.r7_context.source_asset.type	3445	0,345%	1			
json.r7_context.source_asset.rrn	3445	0,345%	32			
json.r7_context.user.rrn	8950	0,895%	30			
json.r7_context.account.type	8950	0,895%	1			
json.r7_context.user.name	8950	0,895%	30			
json.r7_context.user.type	8950	0,895%	1			
json.r7_context.account.rrn	8950	0,895%	30			
json.r7_context.account	1000000	100,000%	0			
json.r7_context.user.domain	524730	52,473%	1			
json.r7_context.user	1000000	100,000%	0			
json.r7_context.account.name	8950	0,895%	31			

Ilustración 4: Metadata del archivo File Access Activity

La estructura (metadata) del archivo crudo (Raw - sin transformaciones) está conformado por las **48 columnas** y **1.000.000 registros** mostradas a continuación, en este se pueden evidenciar cuales son las columnas que cuentan con información nula o vacía su porcentaje y la cantidad de valores distintos (únicos) y su participación con respecto al total de registros lo que puede dar una pista de cuáles podrían ser columnas a omitir o transformar (Ver **Ilustración 4**), visto de otra manera, la cantidad de registros nulos pueden generar sesgo y/o generar uso desbordado de procesamiento que puede ser mitigado (Ver **Ilustración 5**)

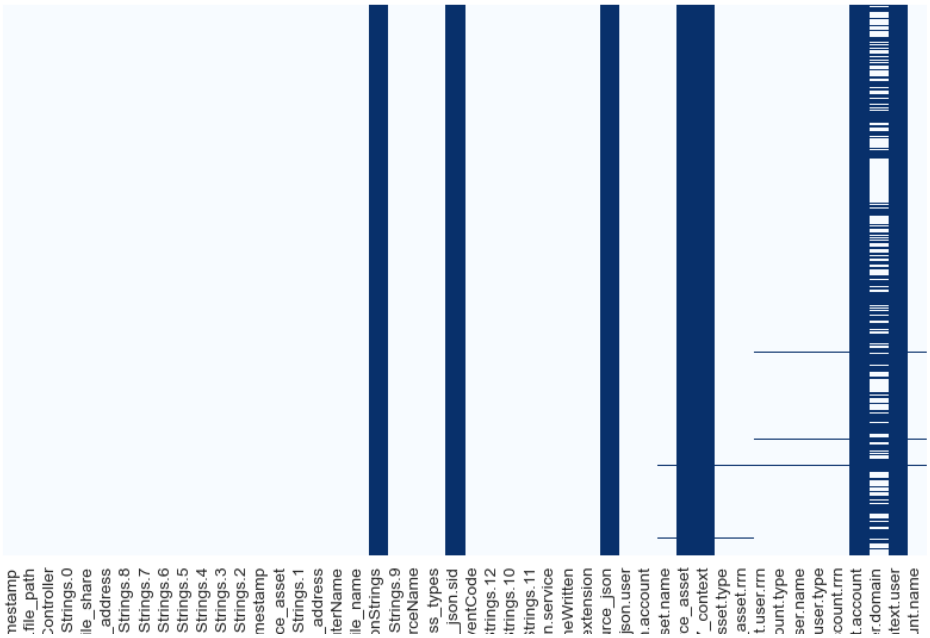


Ilustración 5: Mapa de calor de los valores nulos del archivo File Access Activity

Siendo esta información entregada por el aplicativo Rapid7, y dados los criterios de selección especificados en este mismo documento (Ver Criterio 1 y Criterio 2 de sección Limpieza general de Datos) , se decide omitir columnas utilizando el pipeline desarrollado en lenguaje de programación : Python 3.5 y generando como resultado un total de **25 columnas (Ver Ilustración 6) y 990900 registros** que a su vez fueron procesados para realizar una análisis exploratorio de datos, no se realizaron procesos de imputación de valores por el posible sesgo que esto podría representar , se decidió eliminar los registros de estas columnas desbalanceadas.

Columna	Nulos	Porcentaje de	Valores Distintos		Total Registros	1000000
timestamp	0	0,000%	15888			
json.file_path	0	0,000%	119669			
json.source_json.insertionStrings.0	0	0,000%	55			
json.source_address	0	0,000%	32			
json.source_json.insertionStrings.6	0	0,000%	2928			
json.source_json.insertionStrings.5	0	0,000%	28			
json.source_json.insertionStrings.3	0	0,000%	3919			
json.timestamp	0	0,000%	59767			
json.source_asset	0	0,000%	34			
json.source_json.insertionStrings.1	0	0,000%	57			
json.file_name	0	0,000%	43104			
json.source_json.insertionStrings.9	0	0,000%	119655			
json.source_json.insertionStrings.12	0	0,000%	57			
json.source_json.insertionStrings.10	0	0,000%	34			
json.source_json.insertionStrings.11	0	0,000%	34			
json.source_json.timeWritten	0	0,000%	999858			
json.file_extension	0	0,000%	2433			
json.user	0	0,000%	31			
json.account	0	0,000%	57			
json.r7_context.source_asset.name	3445	0,345%	33			
json.r7_context.source_asset.rrn	3445	0,345%	32			
json.r7_context.user.rrn	8950	0,895%	30			
json.r7_context.user.name	8950	0,895%	30			
json.r7_context.account.rrn	8950	0,895%	30			
json.r7_context.account.name	8950	0,895%	31			

Ilustración 6: Metadata del archivo ajustado File Access Activity

Teniendo ya ajustado los datos, se decide realizar un **análisis exploratorio de datos (EDA),** el cual arrojó los siguientes resultados.

1. Se encontró que de la muestra tomada del mes de abril, los días contaban con un promedio aprox de 190.000 registros y un único contó con 285.125 registros, con el fin de analizar un comportamiento medio, se decide tomar como muestra de análisis el dia 09/04/2024

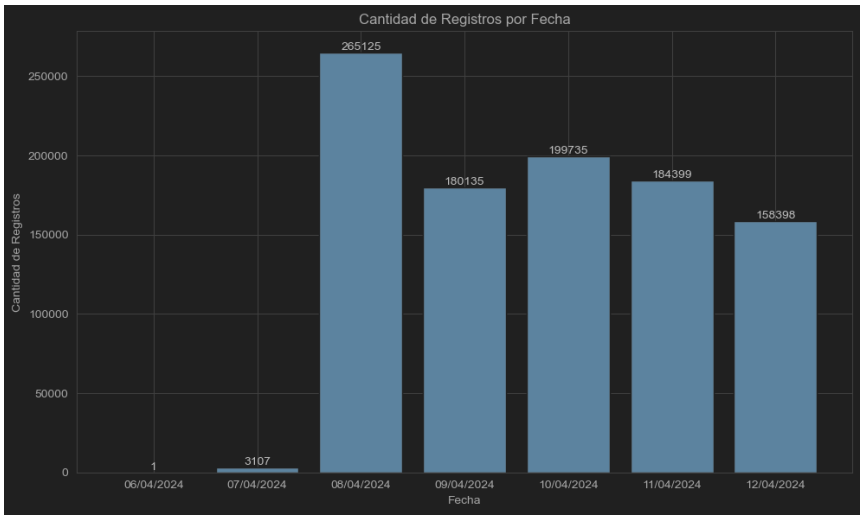


Ilustración 7: Cantidad de registros por dia File Access Activity muestra (no prueba de C2)

2. En la fecha 09/04//2024 se lograron evidenciar valores atípicos (teniendo en cuenta el horario normal de operación 07:00-18:00) , que podrían ser analizados para determinar si estos son o no sospechosos.

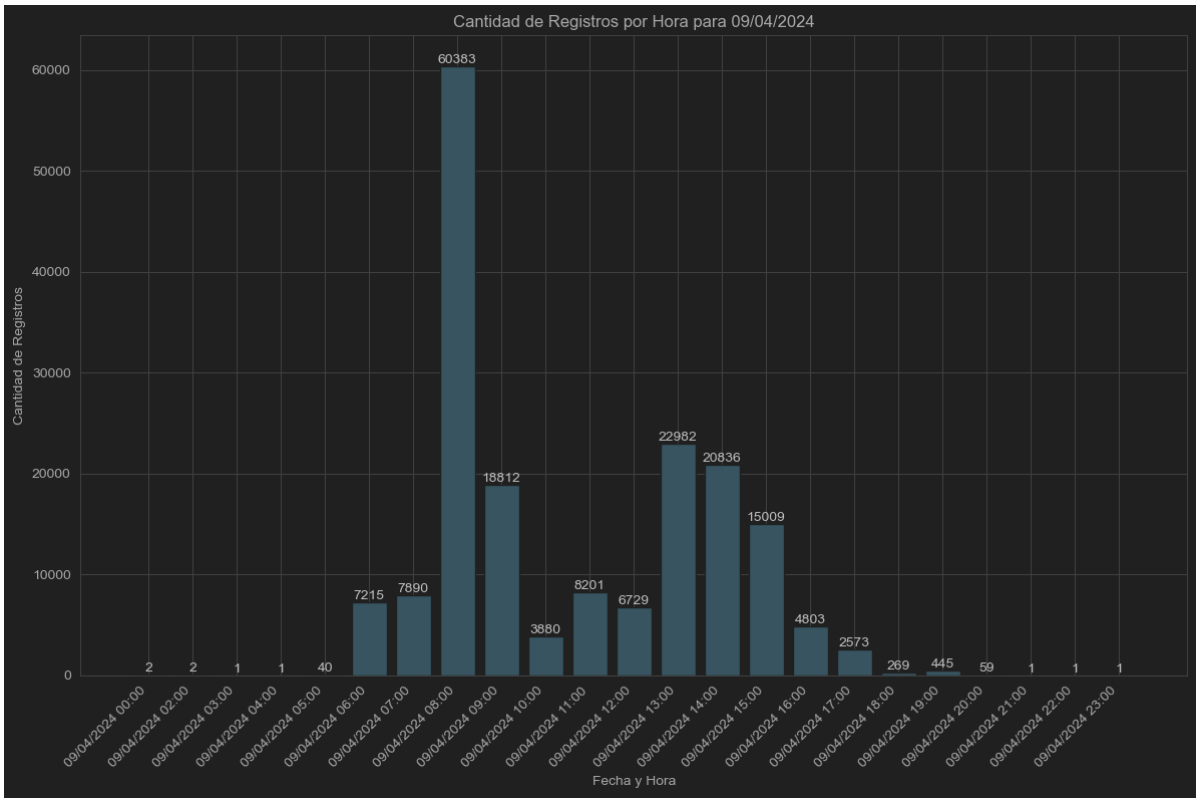


Ilustración 8: Cantidad de registros por hora File Access Activity en el dia 09/04/2024

Una vez realizado el análisis de los registros, se encontró que las operaciones de actividades sobre archivos ejecutadas en estos horarios corresponden a validación y ejecución de políticas (policies) de seguridad programadas de manera automática y ejecutadas por usuarios administradores y también que los archivos que eran accedidos correspondian a ejecutables (.exe) dispuestos en carpetas de políticas de los dispositivos.

Adicionalmente, en la **Ilustración 9** se puede observar el gráfico de correlaciones, en el cual se nota que hay una fuerte relación entre varias de las columnas. Es importante evaluar si debemos eliminar estas variables, ya que tener variables altamente correlacionadas puede traer problemas como la multicolinealidad al momento de elaborar un modelo. Además, mantener ambas variables puede resultar redundante, esto deberá contar un análisis funcional de dichas variables para proceder con la elaboración de modelos.

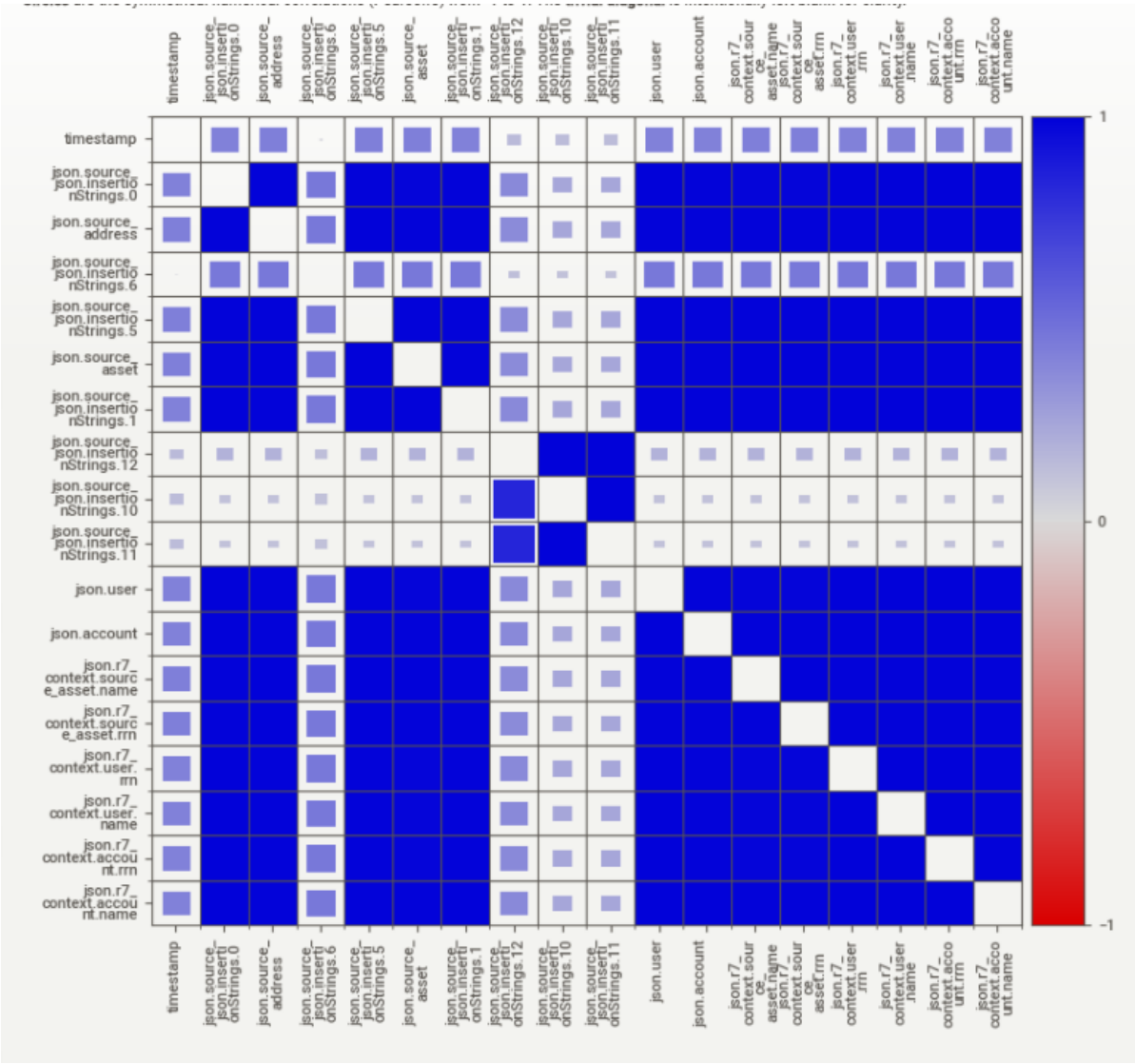


Ilustración 9: Gráfico de correlaciones entre las variables del archivo File Access Activity

El archivo **Firewall Activity** se refiere al monitoreo y registro de eventos y acciones relacionadas con el firewall de red. Esta actividad incluye el tráfico de red que pasa a través del firewall, como conexiones entrantes y salientes, reglas aplicadas, e intentos de conexión bloqueados o permitidos. El firewall actúa como una barrera entre la red interna y externa, controlando y filtrando el tráfico de red según reglas predefinidas.

Al procesar los datos, se identificaron un total de 33 columnas. De estas, 5 columnas presentaban más del 70% de valores nulos, por lo cual fueron eliminadas del análisis. Además, se encontraron 17 columnas con valores nulos que oscilaban entre el 65% y el 5%. Finalmente, se decidió utilizar 16 columnas para el análisis detallado. En las gráficas de datos faltantes (Ilustración 10), se observó una significativa cantidad de valores nulos, especialmente en las columnas con más del 70% de datos

faltantes, las cuales fueron eliminadas para asegurar la calidad del análisis. En la Ilustración 11 se identifican las correlaciones y las relaciones entre distintas variables, ayudando a entender mejor el comportamiento del tráfico de red y la efectividad del firewall en la detección de amenazas.

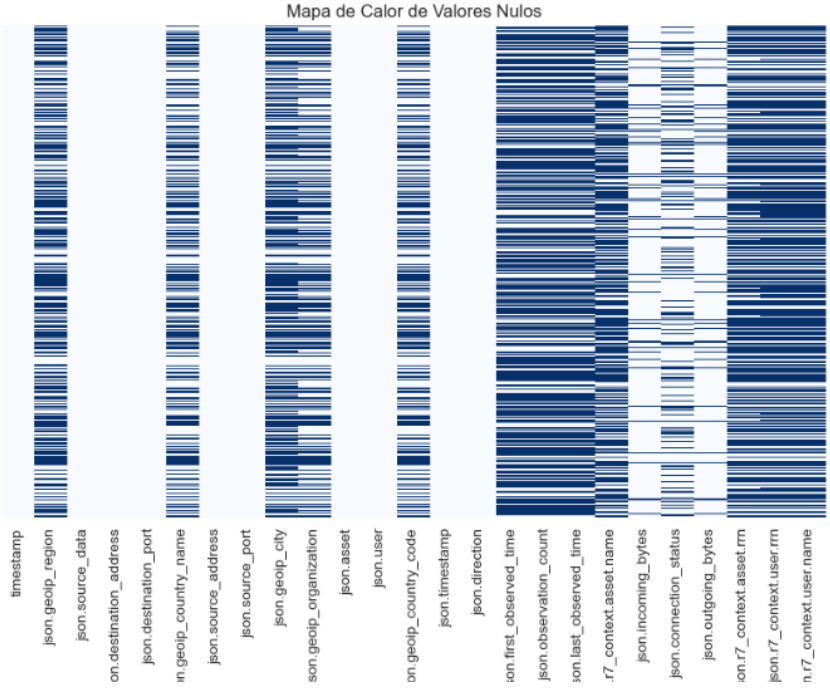


Ilustración 10: Gráfico de valores faltantes del archivo Firewall Activity

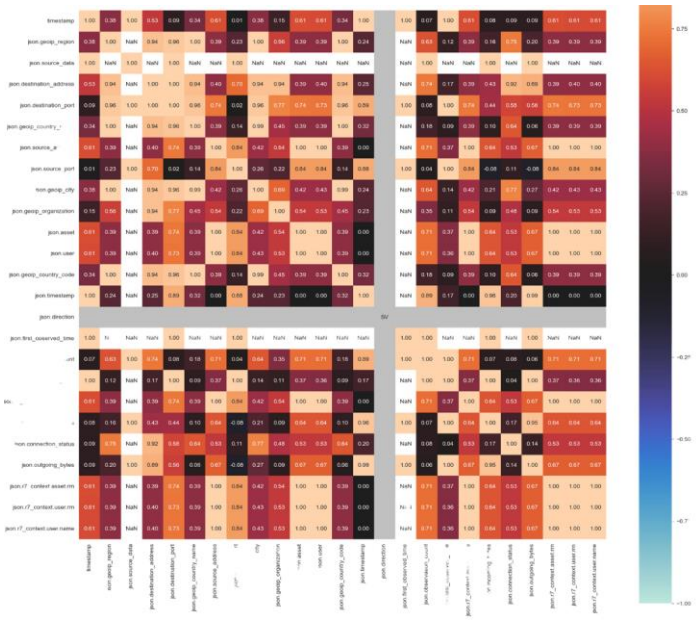


Ilustración 11: Gráfico de correlaciones entre las variables del archivo Firewall Activity

El archivo **IDS Alert** está relacionado con la detección y respuesta a intrusiones en la infraestructura de TI de una organización. Un Sistema de Detección de Intrusiones (IDS) monitorea el tráfico de red o actividades del sistema en busca de comportamientos sospechosos o maliciosos.

En el procesamiento de estos datos, se identificaron 49 columnas inicialmente. Sin embargo, 30 de estas columnas presentaban más del 70% de valores nulos, y por ende, fueron eliminadas del análisis. También se encontraron 6 columnas con valores nulos entre el 65% y el 5%. Al final, se utilizaron 19 columnas para el análisis detallado. La limpieza de datos siguió un proceso similar al del archivo "Firewall Activity". La gráfica de datos faltantes (Ilustración 12) revela un alto porcentaje de valores nulos en múltiples columnas, lo que justificó su eliminación para mantener la integridad del análisis. Analizamos las correlaciones entre las variables (Ilustración 13) para identificar patrones y relaciones relevantes, esenciales para mejorar las estrategias de detección y respuesta a intrusiones.

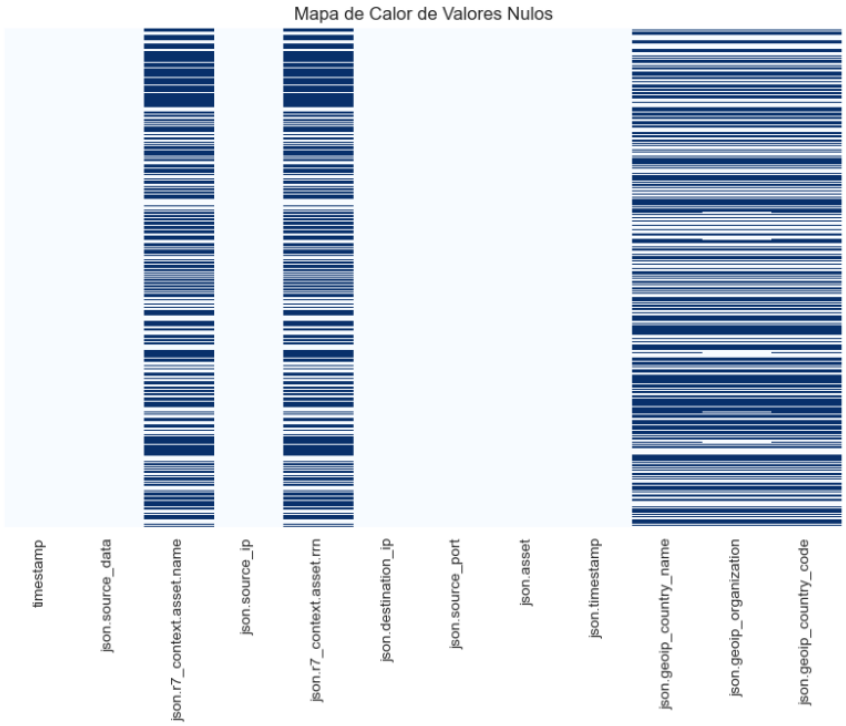


Ilustración 12: Gráfico de valores faltantes del archivo IDS Alert

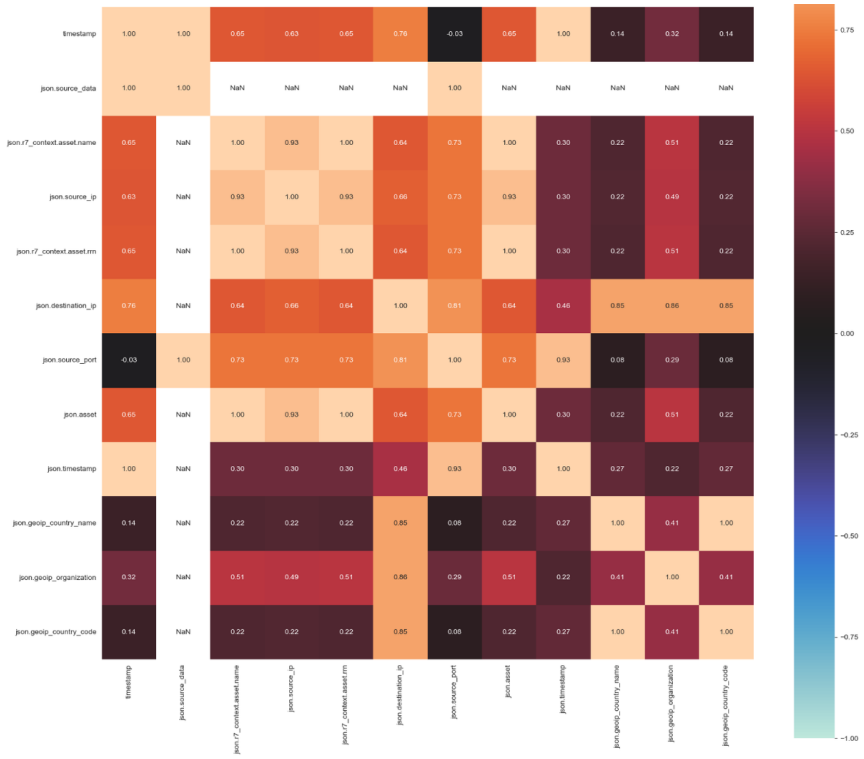


Ilustración 13: Gráfico de correlaciones entre las variables del archivo IDS Alert

El archivo **File Modification Activity** es crucial para la ciberseguridad y administración de sistemas, especialmente en el contexto del COCIB. Este archivo permite detectar modificaciones en archivos dentro de un sistema, lo que puede indicar comportamientos maliciosos o no autorizados. Contiene registros de cambios como la creación, modificación y eliminación de archivos, e incluye datos sobre el usuario responsable, fecha, hora y ubicación de las modificaciones. Este archivo es esencial para la administración de seguridad, proporcionando visibilidad y control sobre los cambios en archivos, permitiendo la identificación de actividades no autorizadas o maliciosas. Facilita la revisión de registros para identificar patrones y anomalías, ayudando a detectar accesos no autorizados, malware y comportamientos sospechosos. Inicialmente, el archivo contenía 47 columnas con 161,248 registros, pero se eliminaron 14 columnas (29.78% del total) debido a criterios de valores nulos. Finalmente, por desbalanceo, agrupación y selección de características, el archivo se redujo a 7 columnas esenciales.

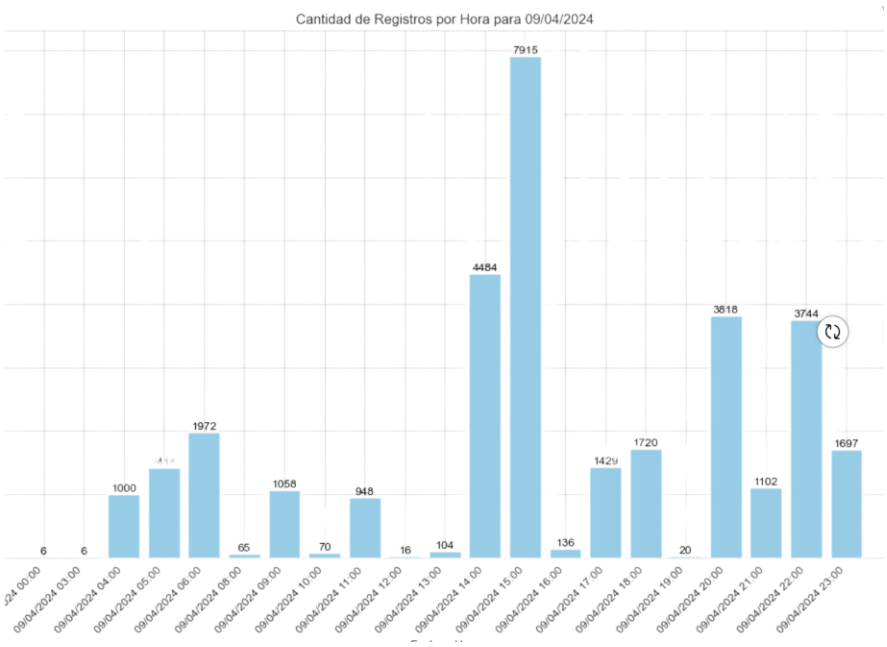


Ilustración 14: Gráfico de distribución de Registros por tiempo del archivo File Modification Activity

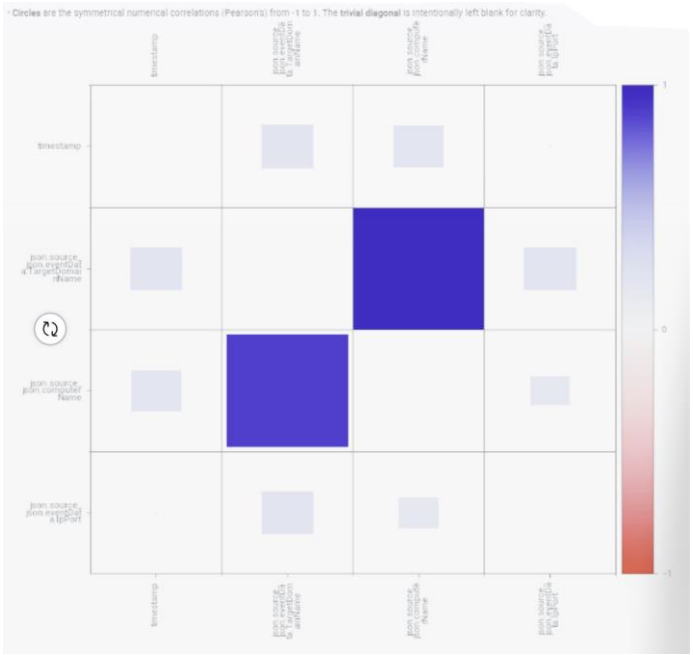


Ilustración 15: Gráfico de correlaciones entre las variables del archivo File Modification Activity

Lecciones Aprendidas: Primer alcance

Al finalizar la primera etapa del proyecto, se han generado aprendizajes y algunas recomendaciones que permitirán potenciar el proyecto en su segunda y siguientes etapas:

1. El entendimiento del negocio: Al igual que con cualquier otra organización o empresa, entender el funcionamiento interno del comando ha sido el primer y mas arriesgado reto en esta fase. La inteligencia militar esta fundamentada en principios que abarcan diferentes tipos de agudeza, como la de combate, la estratégica y la contrainteligencia. En este sentido, la ciberdefensa hace parte de un aspecto relevante en la toma de decisiones, donde los principales objetivos son salvaguardar la armada nacional y, por ende, la seguridad nacional.
2. Entender los datos: Aunque se puede considerar que entender los datos y tablas entregadas por el comando, las cuales son generadas por el Rapid7, haría parte del propio entendimiento del negocio, el análisis de cada una de las tablas con sus respectivos atributos y registros, entender los nulos, faltantes, outliers, el tipo de distribución, entre otros, hace parte de un aprendizaje mucho mayor, donde se pone en juego todo lo aprendido en la presente maestria además de hacer un análisis transversal con la propuesta del COCIB.
3. Decisiones relevantes: Como se mencionó anteriormente, la toma de decisiones en los procesos intermedios permitió el avance en mayor o menor medida del proceso EDA. Las decisiones mas importantes fueron aquellas que involucraron los datos vs el negocio. Un gran ejemplo de ello fue la decisión con los datos Nulos. En un estudio tradicional de ciencia de datos, uno de los aspectos mas importantes es la imputación de datos faltantes. Aunque se consideraron distintas formas de imputar aquellos datos nulos, que en algunas tablas superaba el 50%, lo cierto es que carecemos de conocimiento técnico para imputar, porque hasta este punto, imputar datos de fechas, registros, accesos y ubicación del dominio parece carecer de sentido. En esta toma de decisiones fue fundamental el apoyo de la organización
4. Mirar al futuro: Como se ha descrito en la metodología Asum, el proceso será cíclico en la medida en que tomemos decisiones a futuro, pero que posteriormente se puedan o deban reversar. Todo depende del buen desempeño de los modelos. Las decisiones futuras, la prueba y despliegue de modelos dependerán de un acompañamiento total del comando y de la Universidad.

Recomendaciones

Asi mismo como parte de los aprendizajes, están aquellos desafíos que se encuentran las cuales pueden ser de carácter técnica, profesional o logística. En ese sentido algunas de las recomendaciones que brindamos al comando y al equipo mismo son:

- Se hace necesario contar con repositorios de trabajo conjunto con la entidad de trabajo, esto debe ir acompañado con una gestión estratégica y un gobierno de datos, que permita definir aspectos claves como almacenamiento de volumen grandes de datos (Big Data), gestión estratégica que tenga como centro los datos y su calidad, el custodio de estos, asignación de responsabilidad sobre los datos, privacidad y seguridad de estos. Este proceso debe tener como base la colaboración Comando & Universidad.
- Se hace indispensable tener sesiones de trabajo frecuentes que permitan responder a dudas funcionales de la información recolectada, esta recomendación no solamente va dirigida al comando, es transversal a todo el equipo de trabajo. El trabajo permanente, el cumplimiento en asignación de responsabilidades en los tiempos establecidos y la validación de los tutores expertos y el cliente permitirá un avance eficaz y eficiente.
- Retomar la documentación y en algunos casos, revisar todo el EDA, hará que el equipo reflexiones sobre decisiones tomadas y permitirá reiterarla o redireccionarlas hacía un producto final
- Hacer invitación a demás equipos de trabajo de otras cohortes, además de elegir un equipo interno que sea el encargado de capacitarlos, esto con el ánimo de fortalecer el equipo, aprovechar habilidades y sumar esfuerzos y capacidades en pro de un excelente producto final.

Bibliografía

Armada Nacional de Colombia. (2021). Plan de Desarrollo Naval. *Jefatura de planeación naval*.

Borrero, R. C. (2015). Estado actual de la política pública de ciberseguridad y ciberdefensa en Colombia. *Revista de derecho, telecomunicaciones y tecnología*.

Colombo, H., Sliafertas, M., Pedernera, J., & Kamlofsky, J. (2015). Un Enfoque para Disminuir los Efectos de los Ciber-ataques a las Infraestructuras Críticas. *ResearchGate*.

Consejo Nacional de Política Económica y Social –CONPES . (2020). *Documento 3995. Política Nacional De Confianza Y Seguridad Digital.* Bógota.

Consejo Nacional de Política Económica y Social –CONPES. (2011). *Documento 3701. Lineamientos de Políticas para ciberseguridad y Ciberdefensa. politica de seguridad Nacional* . Bógota: Resolución 3854 de 2009.

Cortina, V. G. (2015). Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario. *Universidad Carlos III de Madrid*.

Crawford, J. (2019). Ciberataque al transporte marítimo. ¿una amenza real o ciencia ficción? *Revista de Marina*, 15-23.

Cujabante, X., Bahamón Jara, M., Prieto Venegas, J., & Quiroga Aguilar, J. (2020). Ciberseguridad y ciberdefensa en Colombia: un posible modelo a seguir en las relaciones cívico-militares. *Revista Científica General José María Córdova, 18*(30).

Departamento Administrativo de la Función Pública. (2022). *Decreto 338 de 2022.* Bógota.

Escuela Superior de Guerra. (2020). Ministerio de defensa nacional comando general fuerzas militares escuela superior de guerra Estrategia Nacionalde Ciberdefensa y Ciberseguridad 2020-2030. *Estrategia Nacional de Ciberdefensa y Ciberseguridad*.

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems,*, 330–347.

Grijalba, P. C. (2020). *Proyecto de actualización de los sistemas Firewall para mejorar la ciberseguridad en la Marina de Guerra del Perú.* Piura: Universidad de Piura.

Hernández, E. G. (2022). Análisis predictivo en Twitter para detectar patrones de personas con tendencia Hacktivista aplicando Big Data, Machine Learning y Deep Learning. *Universidad Cuahtemoc*.

Ministerio de Defensa Nacional. (2021). Disposición 127 de 2021. *Armada Nacional*.

Newmeyer, K. (2015). Ciberespacio, ciberseguridad y ciberguerra. *Escuela superior de guerra naval* , 76-95.

Posada, J. E. (2021). Elementos de ciberseguridad para neutralizar los ataques cibernéticos en las unidades a flote de la armada nacional de colombia. Bógota: Escuela Superior de Guerra General Rafael Reyes Prieto.

Rahul Katarya, & Om , P. (2016). Recent developments in affective recommender systems. *Physica*, 182–190.

Rivero, J. J. (2022). Data science and artificial intelligence: experience in qualitative research. *REVISTA EDUCARE*, 186-201.

Suárez, J. S. (2023). Ciberseguridad: un desafío para las Fuerzas Militares colombianas en la era digital. *Revista Perspectivas en Inteligencia*, 333-359.

Suárez, J. S. (2023). Cybersecurity, a challenge for the Colombian Military in the digital age. *Revista Científica en Ciencias Sociales e Interdisciplinaria, 15*(24), 333-359.