ANALISIS DE DATOS PARA LA CONSTRUCIÓN DEL SISTEMA DE GESTION DE ACTIVOS de INFORMACIÓN DE LA GOBERNACIÓN DEL CAUCA PORJECT CIBER

Con base a las recomendaciones del mentor para hacer el análisis de datos para proyectar un análisis lo más cercano a la realidad se tuvo en cuenta las siguientes fuentes de información:

Dataset secretaria general GOBERNACIÓN DEL ATLÁNTICO:

Propietario de conjunto de datos

Gobernación del Atlántico

Información de la Entidad

Departamento	Atlántico
Municipio	Barranquilla
Nombre de la Entidad	Gobernación del Atlántico
Orden	Territorial
Sector	Función Pública
Área o dependencia	SECRETARIA GENERAL

Información de Datos

Idioma	Español
Cobertura Geográfica	Departamental
Frecuencia de Actualización	Semestral
Fecha Emisión (aaaa-mm-dd)	2020-10-07
Enlace de la fuente	https://www.datos.gov.co/d/8wwe-wqtw

Siguiendo con la problemática de la Secretaría General del Cauca donde se plantea que no se cuenta con arquitectura de ciber seguridad, se recogió análisis de datasets entorno a varios cuenta las siguientes fuentes para aproximarnos a una solución integral para nuestro potencial cliente:

Software malicioso

- Conjunto de datos UNSW-NB15 : este conjunto de datos tiene nueve familias de ataques, a saber, Fuzzers, Análisis, Backdoors, DoS, Exploits, Genéricos, Reconocimiento, Shellcode y Gusanos. Se utilizan las herramientas Argus y Bro-IDS y se desarrollan doce algoritmos para generar un total de 49 características con la etiqueta de clase.
- Conjuntos de entrenamiento de malware : hoy (consulte la fecha de la publicación del blog) los conjuntos de datos clasificados recopilados están compuestos por las siguientes muestras: APT1 292 muestras, Crypto 2024 muestras, Locker 434 muestras, Zeus 2014 muestras
- El conjunto de datos de Drebin : el conjunto de datos contiene 5.560 aplicaciones de 179 familias de malware diferentes. Las muestras se recopilaron entre agosto de 2010 y octubre de 2012 y nos las proporcionó el proyecto MobileSandbox.
- Stratosphere IPS - Capturas de malware, capturas normales, capturas mixtas...
- Desafío de clasificación de malware de Microsoft : se le proporciona un conjunto de archivos de malware conocidos que representan una combinación de 9 familias diferentes. Cada archivo de malware tiene un Id., un valor hash de 20 caracteres que identifica de forma única el archivo, y una Clase, un número entero que representa uno de los 9 nombres de familias.

Aplicaciones web

- Conjuntos de datos de la NSA de West Point : registro de detección de intrusiones de Snort, registros del servicio de nombres de dominio, registros del servidor web, registro agregado del servidor de registros.
- Cargas útiles de ataques web : una colección de cargas útiles de ataques web.
- Cortafuegos de aplicaciones web impulsado por aprendizaje automático : conjunto de consultas buenas y malas a un firewall de aplicaciones web.

URL y nombres de dominio

- Conjunto de datos de URL maliciosas : el conjunto de datos consta de aproximadamente 2,4 millones de URL (ejemplos) y 3,2 millones de funciones.
- cybercrime-tracker : lista de URL maliciosas etiquetadas.
- Lista de dominios de malware - Lista de dominios de malware.
- ZeuS Tracker : ZeuS Tracker rastrea los servidores (hosts) de comando y control de ZeuS en todo el mundo y le proporciona una lista de bloqueo de dominios y de IP.
- URLhaus - URLhaus es un proyecto de abuse.ch cuyo objetivo es compartir URL maliciosas que se utilizan para la distribución de malware.
- StopForumSpam : los datos que se proporcionan aquí representan lo que creemos que solo se utilizará con fines abusivos. Las direcciones IP, los dominios y los

nombres de usuario que se enumeran aquí se mostrarán en los resultados de la API como "incluidos en la lista negra".

Correo electrónico

- Corpus público de spam TREC 2007 : el corpus trec07p contiene 75.419 mensajes: 25.220 de radioaficionado y 50.199 de spam.
- Lista de SPAM - Lista de mensajes de SPAM
- **tarros de miel**
- Colección de conjuntos de datos de DDS : un archivo CSV tar/gzip de una colección de honeypots de AWS. Un archivo CSV zip de dominios y una clasificación de alto nivel de dga o legítimo junto con una subclase de legítimo, cryptolocker, gox o newgoz.
- Threat_Research : repositorio centralizado para volcar datos de investigación de amenazas recopilados de mi red de honeypots.

Suplantación de identidad (phishing)

- Conjunto de datos de sitios web de phishing : en este conjunto de datos, arrojamos luz sobre las características importantes que han demostrado ser sólidas y efectivas para predecir los sitios web de phishing. Además, proponemos algunas características nuevas.

Frecuencia de las contraseñas

- Corpus de frecuencia de contraseñas de Yahoo : este conjunto de datos incluye listas de frecuencia de contraseñas desinfectadas recopiladas de Yahoo en mayo de 2011.

Con el análisis de esta información el equipo pudo determinar el manejo que se debe tener con los activos de información de la Secretaría General de la Gobernación del Cauca de manera detallada. Cada registro representaría un activo de información con atributos como:

- **Contexto:** A qué oficina de la gobernación pertenece.
- **Responsabilidades:** Qué dependencia genera, custodia y tiene acceso a la información.
- **Clasificación:** A qué categoría o programa pertenece.
- **Descripción:** Contenido detallado del activo.
- **Detalles técnicos:** Idioma, formato, medio de conservación.
- **Estado y accesibilidad:** Estado actual, modo de consulta y lugar donde se puede consultar.
- **Ciclo de vida:** Fecha de generación, frecuencia de actualización.
- **Marco legal:** Fundamento constitucional o legal que lo sustenta.

La información obtenida de los dataset infiere que se debe tener en cuenta el desarrollo las siguientes características:

- **Analizar los Metadatos:** Evaluar la necesidad de incluir campos adicionales para enriquecer la descripción de los activos (ej.: palabras clave, resumen ejecutivo).
- **Garantizar la Jerarquización:** Si existen relaciones jerárquicas entre los activos (ej: un documento principal con varios anexos), implementar una estructura de árbol para representarlo.
- **Implementar el Control de Acceso:** Implementar un sistema de roles y permisos para restringir el acceso a la información según el perfil del usuario.
- **Garantizar la preservación Digital:** Establecer políticas y procedimientos para la preservación a largo plazo de los activos digitales.

También el análisis de los tatos nos arrojó que es necesario que se recopile más información, se pueden realizar análisis más profundo que permita detectar y eliminar registros duplicados, Identificar los activos más consultados y los menos utilizados. Como también asegurar la integridad y consistencia de la información y en un proceso 2.0 predecir necesidades futuras utilizando el machine learning para anticipar futuras necesidades de información.

Finalmente se tuvo como resultado del análisis de la bigdata con respecto a Ciberseguridad y teniendo en cuenta la cantidad de información para el análisis de datos, como de malware, información de usuarios y aplicaciones. Implica hacer un análisis dedicado y con mucho más tiempo, sin embargo, para este reto consideramos el siguiente análisis para garantizar la ciber seguridad:

- **Métricas de evaluación:** Seleccionar las métricas de evaluación adecuadas según el problema
- **Conexiones entre dispositivos:** Analizar las relaciones entre diferentes dispositivos y hosts para identificar posibles ataques laterales o propagación de malware.
- **Análisis de tráfico:** Utilizar herramientas de análisis de tráfico de red para investigar paquetes individuales y flujos de tráfico.
- **Detección de anomalías:** Identificar comportamientos inusuales en el tráfico de red que puedan indicar un ataque o una falla del sistema
- **Privacidad:** Asegurar la anonimización de los datos para proteger la privacidad de los usuarios.
- **Sesgos:** Evitar sesgos en los datos y en los modelos.