Actividad 1 Módulo 7

1. ¿Cómo puede ayudar Big Data a enfrentar este problema de correos fraudulentos?

Big Data nos ayuda con la problemática de los correos fraudulentos dado que permite un análisis a gran escala, procesando millones de correos identificando patrones, anomalías y tendencias que con métodos tradicionales no se podría, la detención ocurre en tiempo real a penas se reciben, eliminando amenazas antes que lleguen a los usuarios o afecten el sistema, al utilizar modelos predictivos con machine learning, se pueden entrenar el algoritmo con información histórica, tanto legítimos como fraudulentos, se entrena con estos, los modelos aprenden y mejoran constantemente, considerando además el contenido complejo que podrían tener los correos como texto, metadatos, los enlaces, datos adjuntos, identificando phishing o malware.

2. ¿Qué características del Big Data se aplican aquí?

En este caso, se identifican las 5 V's del Big Data para la detección de correos fraudulentos,

La empresa recibe "millones de correos electrónicos" diariamente, representando una cantidad masiva de datos listos para ser almacenados y procesados (Volumen), los correos llegan de forma continua y en tiempo real, por lo que la solución debe tener la capacidad de analizar este flujo (Velocidad), no son estructurados y provienen de diversas fuentes con diversos formatos (Variedad), al recorrer los mails separa los correos legítimos de los fraudulentos y el *spam* asegurando la calidad (Veracidad) y por último generar valor al proteger a la empresa y a sus usuarios. Al filtrar correos maliciosos, se mejora la seguridad, se previene el fraude, se protege la reputación de la empresa y se optimiza la experiencia del usuario (Valor) .

3. ¿Qué tipo de arquitectura o herramientas se podrían utilizar para implementar una solución eficaz?

Para implementar una solución efectiva, se podría proponer una arquitectura distribuida que combine el procesamiento en tiempo real (*streaming*) y por lotes (*batch*), junto con un conjunto de herramientas especializadas.

Arquitectura Propuesta: Arquitectura Lambda o Kappa

Una **Arquitectura Lambda** sería muy adecuada, ya que permite manejar tanto el análisis en tiempo real como el procesamiento de grandes volúmenes de datos históricos para reentrenar los modelos predictivos a través de:

- **Capa de Ingesta de Datos:**
    - **Herramientas:** Apache Kafka o RabbitMQ. Se usarían para recibir el flujo masivo y constante de correos electrónicos de las diferentes plataformas.
- **Capa de Procesamiento (dividida en dos):**
    - **Capa de Velocidad (Speed Layer) - Tiempo Real:**

Herramientas: Apache Spark Streaming o Apache Flink. Analizarían cada correo en el momento en que llega para una detección inmediata de amenazas conocidas (spam, phishing, malware) aplicando modelos de machine learning ya entrenados.

- **Capa de Lotes (Batch Layer) - Procesamiento por Lotes:**

  Herramientas: Apache Spark o Hadoop MapReduce. Procesarán periódicamente grandes volúmenes de correos almacenados para realizar análisis más profundos, descubrir nuevos patrones de ataque y reentrenar los modelos de Machine Learning con mayor precisión.

- **Capa de Almacenamiento (Storage):**
  - **Herramientas:** Un Data Lake como Hadoop Distributed File System (HDFS) o almacenamiento en la nube (Amazon S3, Google Cloud Storage) para guardar todos los correos.
- **Capa de Servicio y Visualización:**
  - **Herramientas de Machine Learning:** Bibliotecas como Scikit-learn, TensorFlow o Keras (ejecutadas sobre Spark) para crear y entrenar los modelos de clasificación de correos.
  - **Herramientas de Visualización:** Elasticsearch con Kibana o Grafana para que el equipo de seguridad informática pueda monitorear en tiempo real las amenazas detectadas, ver estadísticas y analizar tendencias.

En conclusión, las características clave de la solución son :

- Escalabilidad horizontal para manejar picos de tráfico
- Pipelines de procesamiento paralelo
- Modelos de ML re-entrenables automáticamente
- Integración con sistemas existentes de correo electrónico