Sist, Botter, Beaumont, Ritrovato y Denevi

Caso de Estudio 4: Banco o Fintech
● **Caso de Uso:** Detección de fraude en tiempo real.

● **Descripción Ampliada:** Las entidades financieras enfrentan el reto constante de proteger las transacciones de sus clientes. Un sistema de detección de fraude tradicional es lento y puede fallar con patrones de ataque modernos.
El Big Data lo resuelve así:
○ **Análisis en Tiempo Real:** Cada transacción de tarjeta de crédito, transferencia bancaria o movimiento en una cuenta genera datos que se analizan en milisegundos. El sistema no solo ve la transacción actual, sino que la compara con el historial de compras del cliente (ubicación, hora, monto, tipo de comercio) y con patrones de fraude conocidos a nivel global.
○ **Modelos Predictivos:** Un modelo de machine learning se entrena con millones de transacciones fraudulentas y legítimas para identificar anomalías. Si una compra no encaja con el comportamiento habitual del cliente o con los patrones de la mayoría, el sistema puede bloquear instantáneamente o enviar una alerta.

Objetivos:
● Aplicar una visión integral del ecosistema de Big Data a escenarios reales.
● Conectar los conceptos teóricos (5 V's, almacenamiento, procesamiento, etc.) con sus aplicaciones prácticas.
● Fomentar el pensamiento crítico, la investigación y la colaboración en grupo.
● Desarrollar habilidades de presentación y comunicación de hallazgos.
Instrucciones para los Alumnos:
Formación de Grupos:
○ Utilizar los grupos ya creados en la clase anterior.
Asignación del Caso de Estudio:
○ A cada grupo se le asignará uno de los casos de uso real (se encuentran al final del documento).
Investigación y Discusión:
○ En tu grupo, analicen en profundidad el caso de estudio asignado.
○ Utilicen los siguientes puntos como guía para su análisis y tomen notas de sus respuestas:

1 Las 5 V's del Big Data:
■ Volumen: ¿Qué tipo de datos masivos se generan o manejan? ¿De qué escala hablamos (terabytes, petabytes)?
■ Velocidad: ¿Los datos se procesan en tiempo real o por lotes? ¿Por qué esa velocidad es crucial para el éxito de la empresa?
■ Variedad: ¿Qué tipos de datos se utilizan (estructurados, no estructurados, semi-estructurados)?
■ Veracidad: ¿Qué desafíos de calidad y confiabilidad de datos podrían enfrentar? \
■ Valor: ¿Cuál es el beneficio de negocio (ganancias, eficiencia, satisfacción del cliente) que se obtiene del Big Data en este caso?

2 Almacenamiento:
■ ¿Dónde se almacenarán estos datos? ¿Creen que sería un sistema de archivos distribuido como HDFS, un Data Lake o una base de datos más tradicional?

■ ¿Qué desafíos de escalabilidad y costo enfrentarían al almacenar estos datos?

3 Procesamiento y Análisis:
■ ¿Qué tipo de procesamiento se necesita (por lotes o en streaming)?
■ ¿Qué herramientas de análisis serían las más adecuadas (ej. SQL, Python, machine learning)?

4 Gobernanza y Seguridad:
■ ¿Qué datos sensibles o personales podrían estar manejando? (ej. datos personales de clientes, historial de navegación)?
■ ¿Qué desafíos de seguridad y privacidad tendrían que considerar para proteger la información?

4. Presentación de Hallazgos:
○ Al finalizar el tiempo de discusión, cada grupo presentará un resumen de su análisis al resto de la clase. El objetivo es compartir una visión integral del caso de estudio, conectando todos los puntos analizados.

Respuestas:

1. Las 5 V:

- **Volumen:**
    - En el contexto de un Banco o Fintech, se generan y manejan cantidades masivas de datos transaccionales. Cada transacción de tarjeta de crédito, transferencia bancaria o movimiento en una cuenta genera datos constantemente que nos van a servir para ir generando y creando una base de datos. Además, se considera el historial de compras del cliente, que también contribuye a este volumen.
        - Transacciones financieras (cada operación)
        - Información de clientes (datos personales, documentos, historial)
        - Datos de uso en apps/web (logs, cookies, comportamiento)
        - Datos de cumplimiento legal y regulatorio
        - Imágenes de cheques, DNI, etc.
        - Backups y replicaciones para seguridad
    - Las organizaciones financieras manejan petabytes de datos dependiendo de la escala en la que operan las empresas. La industria bancaria ya en 2009 almacenaba 619 petabytes. Se prevé que el volumen de información digital global alcance los 40 zettabytes (ZB) para 2020, lo que equivale a 5,247 gigabytes por persona. Esto significa que los datos crecen exponencialmente y que incluso lo que hoy parece enorme, será normal en pocos años.

- **Velocidad:**
    - Para la detección de fraude, los datos se procesan en tiempo real (Streaming). Cada transacción se analiza en milisegundos. Esto contrasta

con el procesamiento por lotes (Batch), que es ideal para tareas que no requieren una respuesta inmediata, como reportes diarios.
    - La velocidad es crucial porque el procesamiento de datos debe ser rápido para la toma de decisiones, especialmente en casos sensibles al tiempo como la detección de fraude. El sistema no solo debe ver la transacción actual, sino compararla instantáneamente con el historial del cliente y patrones de fraude globales. La capacidad de procesamiento se mide por la cantidad de datos que se pueden analizar y con qué latencia.

- **Variedad:**
    - Se utilizan diversos tipos de datos. El sistema de detección de fraude compara la transacción actual con el historial de compras del cliente (ubicación, hora, monto, tipo de comercio), que son ejemplos de datos estructurados. Además, se usan patrones de fraude conocidos a nivel global, que pueden derivar de datos no estructurados (como logs, texto de comunicaciones, etc.) o semiestructurados (como logs de servidores).
    - La capacidad de analizar conjuntamente datos estructurados, semiestructurados y no estructurados es clave en Big Data, ya que las fuentes de datos son de cualquier tipo y las bases de datos relacionales tradicionales no pueden manejar toda esta variedad de forma eficiente.

- **Veracidad:**
    - La veracidad se refiere a la calidad, precisión y relevancia de los datos masivos, siendo fundamental para obtener *insights* confiables. En el ámbito financiero, esto es de suma importancia, ya que las decisiones se basan en la fiabilidad de la información.
    - Un desafío clave es asegurar que los datos sean confiables, precisos y actualizados. Para la detección de fraude, la fiabilidad de la información utilizada es crítica. El establecimiento de la fiabilidad de Big Data es un gran reto a medida que la variedad y las fuentes de datos crecen. La calidad de los datos se gestiona con métodos para medir y mejorar su integridad, incluyendo el análisis sintáctico, la estandarización, la validación, la verificación y la coincidencia de registros.

- **Valor:**
    - El beneficio principal del negocio en este caso es la protección de las transacciones de los clientes y la prevención de pérdidas financieras. Al identificar anomalías y bloquear o alertar instantáneamente, el sistema genera valor y confianza en el cliente.

- Las técnicas de Machine Learning/IA se utilizan para encontrar patrones complejos, hacer predicciones y clasificaciones. Los modelos predictivos se entrenan con millones de transacciones fraudulentas y legítimas para identificar anomalías, lo que permite a los bancos detectar fraudes antes de que ocurran. Este enfoque permite a las empresas anticipar la demanda, optimizar productos y marketing, y ser los primeros en adoptar tendencias.

2) Almacenamiento

En un sistema de detección de fraude en tiempo real, se generan y procesan enormes volúmenes de datos de manera constante: transacciones, historiales de usuarios, registros de comportamiento y patrones globales de fraude. Para almacenar esta información de forma eficiente y escalable, se requiere una arquitectura moderna y flexible.

Una opción adecuada es utilizar un sistema de archivos distribuido, como HDFS Hadoop, que permite dividir los datos entre múltiples servidores y replicarlos para asegurar disponibilidad y tolerancia a fallos. También es conveniente incorporar un Data Lake, donde se puedan almacenar datos en crudo, tanto estructurados como no estructurados, sin necesidad de transformarlos previamente. Este enfoque facilita la recopilación de datos diversos provenientes de múltiples fuentes. Además, se pueden usar bases de datos relacionales o NoSQL para registrar los resultados procesados, alertas generadas o información crítica que requiere acceso rápido y frecuente.

El principal desafío en este contexto es la escalabilidad. La infraestructura debe ser capaz de crecer rápidamente para soportar el aumento constante del volumen de datos sin perder rendimiento. A esto se suma el costo del almacenamiento, que obliga a gestionar de forma estratégica qué datos se almacenan, durante cuánto tiempo y en qué tipo de soporte. Para optimizar recursos, es común separar los datos en “calientes” (de acceso frecuente) y “fríos” (consultados ocasionalmente), almacenándolos en sistemas adecuados a sus características.

En resumen, el almacenamiento de un sistema de detección de fraude en tiempo real debe ser robusto, escalable y rentable, combinando diferentes tecnologías según el tipo y uso de los datos.

3. Procesamiento y Análisis

- **Tipo de procesamiento:**
  - **En streaming (tiempo real)**, porque el valor está en identificar y actuar sobre transacciones fraudulentas en milisegundos. Procesar por lotes no sería viable, ya que permitiría que el fraude ocurra antes de detectarlo.
  - En algunos casos, se puede complementar con procesamiento por lotes para entrenar y mejorar modelos predictivos con grandes volúmenes de datos

históricos.

- **Herramientas y tecnologías de análisis:**
    - **Frameworks de streaming:** Apache Kafka, Apache Flink o Spark Streaming, para ingerir y procesar eventos en tiempo real.
    - **Lenguajes y librerías de análisis:** Python (con librerías como Pandas, Scikit-learn, TensorFlow o PyTorch) o R, para el desarrollo y entrenamiento de modelos de machine learning.
    - **Machine Learning y Análisis Predictivo:** modelos supervisados y no supervisados para detección de anomalías, redes neuronales y modelos de clasificación.
    - **Bases de datos en memoria:** como Redis o Memcached, para consultas ultrarrápidas durante el análisis en vivo.

4. Gobernanza y Seguridad

Los datos sensibles que se pueden ver manejados son:

- Informacion personal (nombre, direccion, fecha de nacimiento)
- Datos financieros (CBU, numeros de tarjetas)
- Historial de transacciones
- Datos de autenticacion (usuarios, contraseñas o tokens de acceso)

Luego, ciertos desafios tanto de seguridad como de privacidad serian:

- Fuerte cifrado para evitar robo de datos
- Sistema de inicio y autenticacion robusta
- Adaptacion a las regulacion de seguridad de las distintas localidades
- Implementar sistemas de monitoreo y seguimiento continuo
- Contar con un plan de contingencia en caso de incidentes