Las 5 V's del Big Data

Volumen

- **Tipo de datos:** Transacciones bancarias, compras con tarjeta, transferencias, ubicaciones, horarios, montos, tipo de comercio, dispositivos usados, historial del cliente.
- **Escala:** Miles de millones de transacciones a nivel mundial → fácilmente en **petabytes** si se considera el historial y datos globales.

Velocidad

- **Procesamiento:** En **tiempo real (streaming)**, en milisegundos.
- **Importancia:** Si una transacción sospechosa no se detecta inmediatamente, el fraude puede completarse. La **respuesta instantánea es crítica** para proteger tanto al cliente como a la entidad financiera.

Variedad

- **Tipos de datos:**
    - Estructurados: registros de transacciones, montos, cuentas.
    - Semi-estructurados: logs de eventos, cookies de navegación.
    - No estructurados: posibles comentarios en soporte, correos, etc.

Veracidad

- **Desafíos:**
    - Transacciones legítimas que parecen sospechosas (falsos positivos).
    - Transacciones fraudulentas que pasan desapercibidas (falsos negativos).
    - Datos incompletos o erróneos en el historial del cliente.

- **Solución:** Validación continua del modelo con nuevos datos y feedback de usuarios.

Valor

- **Beneficios para la empresa:**
    - Reducción de pérdidas por fraude.
    - Mayor confianza y satisfacción del cliente.
    - Mejora de la imagen corporativa.
    - Eficiencia operativa (automatización vs análisis manual).

Almacenamiento

¿Dónde se almacenan los datos?

- Lo más adecuado:
    - **Data Lake** para guardar datos diversos y a gran escala.

    - **Sistemas distribuidos** como **HDFS o Amazon S3** para alta disponibilidad y volumen.
    - **Data warehouse**

Desafíos:

- **Escalabilidad:** Los datos crecen constantemente; se necesita una arquitectura flexible.

- **Costo:** Almacenamiento en la nube puede ser costoso si no se optimiza bien.

- **Retención:** ¿Cuánto tiempo conservar los datos históricos? (balance entre valor y costos).

Procesamiento y Análisis

Tipo de procesamiento:

- **Streaming (en tiempo real):** detección instantánea de fraudes en transacciones.
- **Lotes (batch):** para entrenar modelos predictivos con millones de transacciones históricas.

Herramientas adecuadas:

- **Lenguajes:** Python, R, SQL.
- **Plataformas de streaming:** Spark Streaming.
- **Machine Learning**
- **Entornos cloud:** AWS, Azure, GCP.

4. Gobernanza y Seguridad

Datos sensibles:

- Información personal del cliente: nombre, DNI, dirección, historial bancario.
- Datos de tarjetas, IBAN, login a banca online.
- Geolocalización, hábitos de consumo.

Desafíos:

- **Privacidad:** Cumplimiento de leyes como **GDPR** o **Ley de Protección de Datos Personales**.
- **Seguridad:**
    - Encriptación en tránsito y en reposo.
    - Autenticación robusta.
    - Accesos restringidos por roles.
    - Auditorías constantes.

- **Consentimiento del usuario:** Informar sobre el uso de sus datos y obtener su autorización si corresponde.

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 bloquearla instantáneamente o enviar una alerta.