Casos de Estudio

Caso de Estudio 1: Empresa de Streaming (Netflix)

- Caso de Uso: Optimización de la experiencia del usuario y creación de contenido original.
- Descripción Ampliada: Netflix, como líder en streaming, no solo compite con otras plataformas, sino también con el tiempo libre de sus usuarios. Para mantenerlos enganchados, utiliza Big Data para todo:
- Motor de Recomendaciones: Recopila datos de miles de millones de interacciones, incluyendo qué ven los usuarios, qué buscan, a qué hora, qué dispositivos usan, cuándo pausan, rebobinan o abandonan un título. Con esta información, el 80% del contenido que un usuario ve proviene de su motor de recomendaciones.
- Creación de Contenido Original: Antes de producir una serie, Netflix analiza qué actores, directores y géneros son populares en diferentes regiones. Los datos les indican qué historias tienen la mayor probabilidad de ser un éxito, lo que justifica la enorme inversión en producción.
- Optimización del Producto: Mediante pruebas A/B a gran escala, Netflix experimenta con diferentes diseños de interfaz y miniaturas de series para ver qué genera más clics y tiempo de visualización.

1. Las 5 V's del Big Data:

- Volumen: ¿Qué tipo de datos masivos se generan o manejan? ¿De qué escala hablamos (terabytes, petabytes)?
  - Netflix maneja petabytes de información diariamente. Recoge datos de miles de millones de interacciones: historial de reproducción, búsquedas, clics, pausas, rebobinados, abandonos, tipo de dispositivo, ubicación y calidad de conexión.
- Velocidad: ¿Los datos se procesan en tiempo real o por lotes? ¿Por qué esa velocidad es crucial para el éxito de la empresa?
  - Procesa datos en tiempo real y también por lotes.
  - Tiempo real: para mostrar recomendaciones personalizadas instantáneamente o ajustar la calidad del video según la conexión.
  - Por lotes: para analizar tendencias de visualización y planificar producciones originales. La velocidad es crucial porque si la recomendación o la calidad del video no se ajustan en segundos, el usuario puede abandonar la plataforma.
- Variedad: ¿Qué tipos de datos se utilizan (estructurados, no estructurados, semi-estructurados)?
  - Estructurados: datos de suscripciones, facturación, horas vistas.
  - Semi-estructurados: logs de servidores, eventos de reproducción en formato JSON.
  - No estructurados: imágenes, videos, audios y metadatos de contenido.
- Veracidad: ¿Qué desafíos de calidad y confiabilidad de datos podrían enfrentar?
  - Pensamos que netflix podría encontrarse con Desafíos como:
    - Datos incompletos (usuario sin conexión).
    - Información sesgada (usuarios que comparten cuenta).
    - Registros duplicados o errores de captura. Si no controlan la calidad, las recomendaciones pueden ser irrelevantes y dañar la experiencia.
- Valor: ¿Cuál es el beneficio de negocio (ganancias, eficiencia, satisfacción del cliente) que se obtiene del Big Data en este caso?

- Ganancias: contenido más visto y recomendado aumenta retención y reduce cancelaciones.
- Eficiencia: decisiones de producción basadas en datos, evitando gastar millones en series sin demanda.
- Satisfacción: recomendaciones personalizadas y calidad de streaming estable mejoran la experiencia y fidelizan usuarios.

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?
  - Netflix utiliza principalmente un Data Lake en la nube (AWS S3) para guardar datos en su formato original, combinado con sistemas distribuidos como HDFS para procesarlos a gran escala. Las bases de datos relacionales y NoSQL se usan para consultas rápidas y datos transaccionales.

- ¿Qué desafíos de escalabilidad y costo enfrentarían al almacenar estos datos?
  - Escalabilidad: al crecer el número de usuarios y contenido, necesitan añadir capacidad de almacenamiento y procesamiento sin perder velocidad.
  - Costo: almacenar petabytes en la nube es caro; deben optimizar compresión, limpieza de datos y priorizar qué guardar. Un mal manejo puede costar millones extra al año.

3. Procesamiento y Análisis:

- ¿Qué tipo de procesamiento se necesita (por lotes o en streaming)?
  - En streaming: para ajustar calidad de video, detectar problemas de conexión o mostrar recomendaciones instantáneas.
  - Por lotes: para análisis históricos, tendencias de consumo y predicciones de popularidad

- ¿Qué herramientas de análisis serían las más adecuadas (ej. SQL, Python, machine learning)?
  - Creemos que utilizan muchas herramientas de análisis para las diferentes necesidades del negocio y sus diferentes “aspectos”
    - SQL para consultas estructuradas rápidas.
    - Python y R para análisis avanzado y machine learning.
    - Apache Spark y Hadoop para procesamiento masivo distribuido.
    - Modelos de machine learning para predecir gustos y optimizar producción.

4. Gobernanza y Seguridad:

- ¿Qué datos sensibles o personales podrían estar manejando? (ej. datos personales de clientes, historial de navegación)?
  - Datos personales: nombre, correo, método de pago.
  - Historial de visualización.
  - Dirección IP y ubicación geográfica.
  - Dispositivo y sistema operativo usado.

- ¿Qué desafíos de seguridad y privacidad tendrían que considerar para proteger la información?

- Proteger datos contra hackeos o accesos no autorizados.
- Evitar filtraciones de contenido exclusivo antes de su estreno.
- Cifrar datos en tránsito y en reposo.
- Cuidado con las tarjetas cargadas para pagar las suscripciones.