Caso de Estudio: Netflix – Optimización de la experiencia del usuario y creación de contenido original

1. Análisis de las 5 V's del Big Data en Netflix

- **Volumen**:
  Netflix maneja datos de miles de millones de interacciones diarias. Estos datos incluyen comportamiento de visualización, búsquedas, pausas, rebobinados, dispositivos usados, entre otros. La escala puede alcanzar los **petabytes**.

- **Velocidad**:
  Se procesan tanto en tiempo real como por lotes. El sistema de recomendaciones y la interfaz necesitan **respuesta rápida** para mantener el engagement. Por ejemplo, cuando un usuario entra, la plataforma ya debe saber qué sugerirle.

- **Variedad**:
  Utilizan datos estructurados (como base de usuarios), semi-estructurados (logs de navegación) y no estructurados (comentarios, imágenes de miniaturas, metadatos de películas, etc.).

- **Veracidad**:
  Los desafíos están en asegurar que los datos reflejen correctamente las preferencias del usuario, evitando sesgos (por ejemplo, que un usuario vea algo por error o deje algo pausado sin intención real). Se requiere **limpieza y validación constante**.

- **Valor**:
  El uso de Big Data le permite a Netflix ofrecer recomendaciones personalizadas (el 80% de lo que se ve proviene de ellas), crear contenido original con mayor probabilidad de éxito (House of Cards fue un ejemplo pionero), y mejorar la interfaz, lo que se traduce en **mayor retención, fidelización y rentabilidad**.

2. Almacenamiento

- Es probable que utilicen un sistema híbrido, incluyendo:

  - **Data Lakes** para almacenar datos sin estructura específica.
  - **Sistemas distribuidos** como **HDFS** o almacenamiento en la nube (ej. Amazon S3) para escalar de forma horizontal y económica.

- **Desafíos**:

  - Escalabilidad: los datos crecen constantemente y deben ser accesibles rápidamente.
  - Costo: manejar almacenamiento eficiente diferenciando datos "calientes" (muy usados) de "fríos" (consultados esporádicamente).

3. Procesamiento y Análisis

- **Procesamiento**:
    - Se necesita **procesamiento en streaming** para personalización en tiempo real.
    - También usan **procesamiento por lotes** para análisis más profundos (preferencias por región, análisis de tendencias de largo plazo).

- **Herramientas**:
    - **Apache Spark** para procesamiento en memoria.
    - **Machine Learning** para generar modelos predictivos (recomendaciones, éxito de contenido).
    - **Python, SQL y herramientas de visualización** para análisis exploratorios y dashboards internos.

4. Gobernanza y Seguridad

- **Datos sensibles**:
    - Datos personales como nombre, edad, ubicación, historial de visualización, preferencias.
    - Potencial uso de datos de comportamiento para segmentación.

- **Desafíos**:
    - Cumplir con regulaciones como **GDPR** (en Europa) y leyes locales de protección de datos.
    - Asegurar la **autenticación** y **autorización** en el acceso a la información.
    - Aplicar **encriptación en tránsito y en reposo**.