Caso de Estudio 5 – Sector Salud y Farmacéutico

- **Volumen:** Se manejan enormes cantidades de datos como secuencias completas de ADN, historiales médicos extensos, datos de sensores y dispositivos médicos, redes sociales y movilidad poblacional. Escala: terabytes a petabytes.

- **Velocidad:** En monitoreo de pandemias, el procesamiento en tiempo real permite alertas tempranas de brotes, control de disponibilidad hospitalaria y análisis de movilidad. La rapidez es clave para salvar vidas.

- **Variedad:** Datos estructurados (registros médicos, bases genéticas), no estructurados (redes sociales, imágenes médicas) y semi-estructurados (sensores, formularios).

- **Veracidad:** Retos como datos incompletos o erróneos, fuentes poco confiables y diferencias de formato entre instituciones.

- **Valor:** Desarrollo de fármacos personalizados más efectivos, predicción de enfermedades, monitoreo rápido de pandemias y mayor eficiencia en la investigación médica.

Almacenamiento: Lo ideal es un Data Lake o sistema distribuido (HDFS) para integrar datos de distintas fuentes. Desafíos: escalabilidad y alto costo de almacenamiento a gran escala.

Procesamiento y análisis: Streaming para datos en tiempo real y procesamiento por lotes para estudios genéticos o desarrollo de medicamentos. Herramientas: Python, R, machine learning (TensorFlow, Scikit-learn), SQL y Apache Spark/Hadoop.

Gobernanza y seguridad: Datos sensibles como información personal, genética e historiales clínicos. Desafíos: cumplir leyes de protección de datos, prevenir hackeos, anonimizar información y restringir accesos.

    - .