Caso de Estudio 5: Sector de la Salud y Farmacéutico

- Caso de Uso: Investigación médica y desarrollo de nuevos medicamentos.
- Descripción Ampliada: El Big Data está revolucionando la medicina, acelerando la investigación y mejorando los tratamientos.
    - Genómica y Datos Clínicos: Se analizan grandes volúmenes de datos genómicos (la secuencia completa del ADN de miles de personas) junto con sus historiales médicos, datos de dispositivos médicos y estilos de vida. Esto permite encontrar correlaciones entre genes y enfermedades, lo que antes era imposible de hacer a gran escala.
    - Fármacos Personalizados: Con estos análisis, se pueden desarrollar medicamentos que se adapten al perfil genético de un paciente, aumentando su efectividad y reduciendo los efectos secundarios.
    - Monitoreo de Pandemias: Durante la pandemia de COVID-19, se utilizaron datos masivos de movilidad, redes sociales, información de hospitales y pruebas de laboratorio para predecir la propagación del virus y tomar decisiones de salud pública.

1. Las 5 V's del Big Data:

- Volumen: ¿Qué tipo de datos masivos se generan o manejan? ¿De qué escala hablamos (terabytes, petabytes)?
    - Se recopilan datos masivos respecto a la información genética de las personas (secuencia completa del ADN), historial médico, datos de dispositivos médicos, estilo de vida, y a su vez, se almacena información sobre enfermedades estudiadas, para relacionar la información de cada individuo con posibles afecciones que estos puedan presentar actualmente o en el futuro, realizando predicciones sobre su estado de salud y su evolución en el tiempo, relacionando los medicamentos pertinentes. Hablamos de una escala de datos en terabytes.
- Velocidad: ¿Los datos se procesan en tiempo real o por lotes? ¿Por qué esa velocidad es crucial para el éxito de la empresa?
    - Consideramos que la información se procesa por lotes y por streaming, dependiendo el caso de uso y la urgencia de la información.

    - Si se trata de un diagnóstico donde debemos analizar si una persona está enferma o no según síntomas o análisis de imágenes, o si se debe recomendar un medicamento en particular según los datos que ingresan de un paciente, la información debe procesarse en tiempo real.
    - Por otro lado, si se utilizan los datos con un fin más exploratorio y de investigación por parte de la industria farmacéutica, esto puede realizarse en lotes. En el caso del procesamiento en tiempo real, la velocidad es crucial, ya que se necesita un diagnóstico veloz, preciso y correcto para poder recomendar y asignar el medicamento o vacuna correcta para la persona que esté sufriendo algún síntoma en particular.
    - También, en el caso del COVID, el monitoreo constante en tiempo real de las consecuencias del virus y cómo esto fue afectando a la población era crucial, ya que se debían tomar medidas de salud pública acertadas y correctas para evitar una crisis masiva, tanto de salud pública como política.
- Variedad: ¿Qué tipos de datos se utilizan (estructurados, no estructurados, semi-estructurados)?
    - Se utilizan datos estructurados, como tablas de datos que contienen información específica sobre los pacientes como datos demográficos, genéticos, historiales, hobbies, estilo de vida.
    - Por otro lado, también pueden utilizarse datos no estructurados, como archivos de audio, imágenes o videos, provenientes de sitios web, redes sociales, recetas u órdenes manuscritas por doctores, etc.
    - Por último, también se utilizan datos semi-estructurados, como registros de inicio de sesión en sitios web de portales de pacientes, o datos de sensores o máquinas relacionados a los dispositivos médicos.
- Veracidad: ¿Qué desafíos de calidad y confiabilidad de datos podrían enfrentar?
    - Falta de confianza en la información, ya que se relaciona con temas delicados de salud y toma de decisiones en base al bienestar de las personas.

    - Calidad de datos: datos poco precisos, que sean ingresados de forma incorrecta o estén anotados con errores.
    - Variedad de datos clínicos sin digitalizar
    - Protección y privacidad de los datos personales de los individuos
- Valor: ¿Cuál es el beneficio de negocio (ganancias, eficiencia, satisfacción del cliente) que se obtiene del Big Data en este caso?
    - El valor de la big data en un caso como el de la salud y la farmacéutica es único, ya que brinda información precisa y relacionada a casos reales sobre enfermedades, diagnósticos, síntomas, genes, y posibles afecciones. Puede ayudar a prevenir una enfermedad de un paciente, o indicarle la medicación precisa para su mejora, o brindar la información necesaria para la toma de decisiones ante una situación particular, como puede ser una crisis mundial de salud, como lo fue la pandemia del COVID-19.
    - Se relaciona principalmente a la eficiencia de la industria de la salud y farmacéutica, su trabajo en conjunto, y su contribución al bienestar general. A su vez, a mejor performance, la investigación relacionada a la predicción de nuevas afecciones y nuevos medicamentos, puede aumentar las ganancias de la industria farmacéutica. Por último, el seguimiento de los individuos, para poder prevenir enfermedades en su vida futura, y el asesoramiento certero y preciso basado en información histórica, puede ser crucial para tener una vida saludable y cuidada, siguiendo las recomendaciones recibidas.

2. Almacenamiento:

- ¿Dónde se almacenarán estos datos?
    - Almacenaremos la información en la nube que nos sea sencillo para poder actualizar los datos sobre los pacientes, que permite acceso remoto y escalabilidad flexible. Facilitando la integración de nueva información clínica, genómica y de estilo de vida de los pacientes en tiempo real.
- ¿Creen que sería un sistema de archivos distribuido como HDFS, un Data Lake o una base de datos más tradicional?

    - Sería un sistema HDFS ya que permite manejar grandes volúmenes de datos heterogéneos y mantener la confidencialidad de la información médica.
- ¿Qué desafíos de escalabilidad y costo enfrentarían al almacenar estos datos?
    - La combinación de historiales médicos, datos genómicos, imágenes y datos de estilo de vida implica un aumento constante y acelerado del almacenamiento requerido.
    - Se necesitan sistemas de procesamiento que permitan consultas rápidas y análisis inmediatos, especialmente para investigación y desarrollo de medicamentos personalizados.
    - Para el caso de salud se deben conservar los datos por varios años, lo que implica costos de almacenamiento sostenidos.
    - La presencia de imágenes médicas, secuencias genómicas y registros complejos requiere herramientas de análisis como Python

3. Procesamiento y Análisis:

- ¿Qué tipo de procesamiento se necesita (por lotes o en streaming)?
    - Se necesitan ambos procesamientos:
    1. por lotes: Para el análisis genómico y de datos clínicos históricos, donde se trabaja con grandes volúmenes de datos acumulados
    2. en streaming: para el monitoreo de pandemias y para el seguimiento en tiempo real de virus, como el covid.
- ¿Qué herramientas de análisis serían las más adecuadas (ej. SQL, Python, machine learning)?

    Algunas herramientas que serían adecuadas serían:
    - SQL: que permite gestionar y analizar grandes bases de datos, como por ejemplo datos clínicos. Ya que se analizan grandes volúmenes de datos, y también se buscan correlaciones entre ellos, no podemos dejar afuera esta herramienta.
    - Python: ya que a la vez, permite una mejor visualización de los datos. También hacer gráficos y crear modelos que ayuden a entenderlos mejor.

    - Machine learning: sirve para encontrar patrones en los datos y predecir cosas, como cómo un medicamento puede funcionar mejor en cada persona. De acuerdo al perfil de cada paciente, machine learning sería una buena herramienta para lograr un diagnóstico más personalizado.

4. Gobernanza y Seguridad:

- ¿Qué datos sensibles o personales podrían estar manejando? (ej. datos personales de clientes, historial de navegación)?
    - Los datos que se están analizando tienen que ver con información personal y de identificación de cada paciente, datos de salud como diagnóstico médico, antecedentes, adn, y en cuanto a su estilo de vida como que comen, si hacen actividad física, como descansan, etc.
    - También datos con respecto a la medicación que consumen.
- ¿Qué desafíos de seguridad y privacidad tendrían que considerar para proteger la información?
    - Se podría enmascarar y encriptar la información más personal de los pacientes
    - implementar controles de acceso, donde solo el personal autorizado pueda ver o modificar la información, y utilizar autenticación para aumentar la seguridad al ingresar a los sistemas.
    - Otro desafío es garantizar que durante la transmisión de datos, por ejemplo entre hospitales o laboratorios, la información esté protegida con protocolos seguros, evitando que se manipule.