1. Las 5 V's del Big Data:

■ Volumen: ¿Qué tipo de datos masivos se generan o manejan? ¿De qué escala hablamos (terabytes, petabytes)?

- Amazon maneja datos masivos que recibe de millones de transacciones a diario, de los historiales de compras, búsquedas, reseñas y etc. La escala llega/ puede llegar a ser de PetaBytes.

■ Velocidad: ¿Los datos se procesan en tiempo real o por lotes? ¿Por qué esa velocidad es crucial para el éxito de la empresa?

- Los datos procesados por amazon son a tiempo real. Esto es porque tienen que mostrar recomendaciones en el momento y hacer actualizaciones de si hay/no hay stock en el momento. Sin la velocidad de tiempo real amazon podría vender productos que no tiene en stock (basicamente mas tiempo de entrega, o cancelación de compra así generando insatisfacción en el cliente)

■ Variedad: ¿Qué tipos de datos se utilizan (estructurados, no estructurados, semi-estructurados)?

- Datos Estructurados como ventas o inventarios, Semiestructurados como clics y No Estructurados como imágenes o videos.

■ Veracidad: ¿Qué desafíos de calidad y confiabilidad de datos podrían enfrentar?

- Su principal desafío es la calidad/confiabilidad de las reseñas y de los datos desactualizados o mal informados sobre stock. Una reseña falsa puede afectar (bien como mal) para la decisión de un cliente y el mal manejo de stock cambia el flujo del ecommerce.

■ Valor: ¿Cuál es el beneficio de negocio (ganancias, eficiencia, satisfacción del cliente) que se obtiene del Big Data en este caso?

- El uso de big data ayuda a Amazon a mejorar la experiencia del cliente teniendo recomendaciones adecuadas a cada usuario, aumenta las ventas y puede llegar a reducir costos de logística.

2.Almacenamiento:

■Podría utilizar un Data Lake que puede estar en la nube, por ejemplo Amazon s3 y funcionar sobre hdfs

■Un desafío puede ser escalar la infraestructura a nivel global, administrar el costo del almacenamiento masivo, guardar lo que es útil y no estar guardando datos innecesarios.

3. Procesamiento y Análisis:

■ ¿Qué tipo de procesamiento se necesita (por lotes o en streaming)?

- Para este caso, se puede llegar a necesitar ambos. Esto se debe a que podríamos utilizar un procesamiento de streaming (básicamente tiempo real) para las recomendaciones y actualización de stock. Pero esto, se puede complementar en un procesamiento de lotes para analizar tendencias y así ayudar con la actualización de stock. Un ejemplo fácil es las tendencias por estaciones, y así manejar el stock verano-invierno correctamente.

■ ¿Qué herramientas de análisis serían las más adecuadas (ej. SQL, Python, machine learning)?

- Apache Spark , para datos rápidos en memoria.
- Mucha machine learning ya que para las recomendaciones para los usuarios se necesita saber de sus gustos, clics, impresiones ,etc.
- SQL y/o python → conllevan a un buen análisis sobre los datos.
- En el caso de que se quiera , se puede usar power bi para mostrar los datos.

4. Gobernanza y Seguridad:
■Estarán lidiando con datos sensibles como datos personales de los consumidores (dirección, tarjeta de crédito, historial de compras, dni y etc.).
■ Los desafíos serían detectar estos datos que son sensibles y poder encriptarlos para que estén más seguros.