

**1. Define qué es Big Data y menciona al menos tres características clave.**

Conjunto de datos extremadamente grande y difícil de procesar. Creados a grandes velocidades y constantes. Las herramientas de datos normales por llamarlos así presentan dificultades al procesar el Big Data.

**2. ¿Cuáles son las tres "V" originales de Big Data? Explícalas brevemente.**

Velocidad con la que se procesan los datos ; variedad de los mismos procedentes de diferentes fuentes como audio, video, texto ; volumen del almacenamiento.

**3. Enumera al menos cinco fuentes de datos que puedan contribuir al volumen de datos en un entorno de Big Data.**

Redes Sociales,

Plataformas de Streaming como Max o Netflix,

Transacciones como las realizadas en Amazon

Registros de geográficos recorridos por los automóviles

Información registrada por los 'wearables'

**4. Explica la diferencia entre datos estructurados, no estructurados y semiestructurados en el contexto de Big Data.**

La diferencia reside en la manera en que son almacenados, como en tablas o registros. Los datos estructurados son fácilmente identificables como lo son las tablas relaciones, por citar un ejemplo. Los semiestructurados como el mismo nombre puede referir, que no presentan una estructura perfectamente definida como los datos estructurados, pero sí presentan una organización definida en sus metadatos donde describen los objetos y sus relaciones. Los no estructurados pueden entenderse como los archivos de audio y video, texto e imágenes.

**5. ¿Por qué el procesamiento en tiempo real es un desafío importante en Big Data? Proporciona un ejemplo.**

El análisis de datos en tiempo real requiere de estructuras diferentes para el procesamiento. Los métodos tradicionales como una hoja de cálculo son insuficientes. No sólo eso, el volumen de datos requiere ser almacenado en servidores cada vez más grandes

**6. ¿Qué son los modelos sintéticos en Big Data y para que se utilizan?**

Se utilizan cuando los datos presentados son insuficientes para el análisis. Los datos sintéticos funcionan como un lego para complementar el análisis. Estos son generados por técnicas de ciencia de datos.

**7. Describe brevemente la importancia de la variedad de datos en Big Data y cómo puede afectar el análisis de datos.**

Los datos nos ayudan a encontrar caminos más fáciles para la toma de decisiones, la variedad de los mismos puede complementar análisis previamente presentados. Cómo se ven afectados depende del método para tratarlos. Los datos no estructurados requieren tratamientos diferentes que aquellos estructurados.

**8. ¿Cuál es el papel de Hadoop en el procesamiento y almacenamiento de Big Data? Menciona al menos dos de sus componentes principales.**

La capacidad de Hadoop para dividir la data en fragmentos pequeños y hacer el análisis más rápido. Resilientes, Distribuidos en datasets.

**9. ¿Qué significa el término "veracidad" en el contexto de Big Data y por qué es importante?**

La calidad de los datos. El término 'Garbage in, Garbage Out' refiere a que aquello que se utiliza como input de algún modelo dado, determinará la calidad del output. Es importante que a través de métodos de ciencia de datos se use data con calidad.

**10. ¿Cuáles son algunos de los desafíos de seguridad y privacidad que se enfrentan en el manejo de Big Data? Proporciona ejemplos.**

El uso de datos personales debe ser tratado con suma cautela. El escándalo de Facebook y Cambridge Analytica del 2011 representa un claro ejemplo de la privacidad de los datos. Datos usados sin consentimiento del usuario para ser empleados en campañas. La seguridad, por ejemplo, el caso Ashley Madison: hackeo masivo que expuso los datos de millones de personas casadas infieles.

**11. Explica cómo se pueden utilizar las redes sociales en el análisis de Big Data y menciona al menos dos tipos de análisis que se pueden realizar.**

Las redes sociales han sido en los últimos tiempos crítica de laboratorios gigantes de comportamiento. Dada la interacción del usuario con páginas ad hoc de sus preferencias como políticas, sociales, deportivas representan para las compañías

información trascendental. El análisis de sentimiento, por ejemplo, permite a las empresas conocer la opinión de usuarios respecto a post publicados. Aunque criticado, el análisis de las palabras oídas por el celular para arrojar publicidad de las palabras escuchadas.

**12. ¿Qué es el aprendizaje automático (machine learning) y cómo se aplica en el análisis de Big Data?**

Refiere al aprendizaje máquina, donde a través de algoritmos se es capaz de identificar patrones en la data. Como lo puede ser la detección de ciertos tipos de cánceres a través de la constante alimentación con hardcode a la máquina.

**13. ¿Cuáles son las ventajas y desventajas del almacenamiento distribuido en Big Data?**

Una ventaja es la velocidad de procesamiento de la data al dividirla en partes pequeñas- Una desventaja es el costo . Configurar los servidores requiere habilidades especializadas y herramientas avanzadas. La latencia de los nodos de distribución puede variar.

**14. ¿Por qué la escalabilidad es un aspecto crucial en los sistemas de Big Data?**

Permite agregar más nodos para manejar volúmenes crecientes de datos sin problemas significativos en el rendimiento.

**15. Explica la importancia de la visualización de datos en el análisis de Big Data. Menciona al menos dos tipos de visualización.**

Hay un ejemplo dado en el curso Data Science impartido por IBM. Refiere que la visualización de datos es como ir en la carretera, dar la vuelta detrás de la montaña y asombrarse con el paisaje. Decir ¡wao! Es permitir a la audiencia asombrarse con la información encontrada. Puede encontrarse data visualization como treemaps wordclouds

**16. ¿Qué es la variabilidad en Big Data y cómo puede afectar el análisis de datos?**

Los diferentes tipos de datos como estructurados, no estructurados o semiestructurados. Las herramientas empleadas para el tratamiento no serán las mismas. SQL no podrá, por ejemplo, tratar con datos no estructurados.

**17. Describe dos aplicaciones prácticas de Big Data en la industria o la**

## **investigación.**

El análisis de imágenes en campos de cultivo para mejorar los procesos de siembra o crecimiento de las plantas. El uso de IAs para dibujar la estructura de las proteínas.

### **18. ¿Qué significa el término "valor" en el contexto de Big Data y cómo se puede obtener valor a partir de los datos?**

El valor puede referir al beneficio o utilidad resultado del análisis. Dependerá del contexto donde los datos son recopilados y empleados para determinar la utilidad. En el mundo empresarial, por ejemplo, la segregación de clientes a través del pago a empresas como Meta permite encontrar masivamente clientes potenciales.

### **19. ¿Cuál es el papel de Spark en el procesamiento de Big Data y que ventajas ofrece en comparación con otras tecnologías?**

Su papel yace en la capacidad de tratar volúmenes masivos de datos. Es fácil de usar, permite el procesamiento de datos en tiempo real y es compatible con modelos de machine learning.

### **20. Enumera tres desafíos éticos y legales relacionados con el uso de Big Data en la toma de decisiones.**

Muchos modelos de machine learning son catalogados de racistas al arrojar resultados que afecten ciertos grupos. La presentación de los datos arrojados por los modelos, debe ser tratado con cautela y profesionalismo.

Muchas de las prácticas de recolección de datos se siguen haciendo sin consentimiento del usuario o usando técnicas de engaño para recolectar la data.