

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

Big Data se refiere al manejo y análisis de conjuntos de datos extremadamente grandes y complejos que no pueden procesarse con herramientas tradicionales.

Características clave:

- Volumen:
- Variedad
- Velocidad

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

- Volumen: Cantidad de datos generados, que requiere soluciones avanzadas de almacenamiento.
- Variedad: Diversidad de fuentes y formatos (texto, imágenes, videos, etc.).
- Velocidad: Rapidez con la que los datos se producen y necesitan ser procesados.

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

- Redes sociales.
- Sensores.
- Bancos
- Dispositivos móviles.
- Datos de navegación web.

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

- Estructurados: Datos organizados en tablas (bases de datos relacionales).
- No estructurados: Información sin formato definido (imágenes, videos).
- Semi-estructurados: Datos con una organización parcial (JSON, XML).

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

Porque se produce una cantidad enorme de datos en muy poco tiempo, lo que vuelve complicado procesarlos todos en poco tiempo. Por ejemplo, las transacciones bancarias.

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

Son datos generados artificialmente para simular escenarios reales y probar sistemas. Se usan para entrenar modelos de machine learning.

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

Al combinar múltiples fuentes de datos es posible obtener una visión más completa de un fenómeno. Por ejemplo, en el comercio, analizar comentarios en redes sociales junto con datos de inventarios permite entender tanto las tendencias de los consumidores como el rendimiento de los productos.

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

Permite el procesamiento distribuido de grandes conjuntos de datos en grupos de computadoras utilizando modelos de programación simples. Está diseñado para escalar desde servidores únicos hasta miles de máquinas, cada una de las cuales ofrece computación y almacenamiento local. En lugar de depender del hardware para ofrecer alta disponibilidad, la biblioteca en sí está diseñada para detectar y manejar fallas en la capa de aplicación, entregando así un servicio de alta disponibilidad sobre un grupo de computadoras, cada una de las cuales puede ser propensa a fallas.

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

La veracidad se refiere a la calidad y fiabilidad de los datos. Es crucial para garantizar análisis precisos.

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

- Protección de datos personales
- Prevención de brechas de seguridad
- Ética en el uso de los datos

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.

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

Es una rama de la inteligencia artificial que permite a las máquinas aprender patrones de datos sin programación explícita.

Se aplica en identificación de fraudes, recomendaciones personalizadas, predicción de ventas.

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

Ventajas:

- Escalabilidad.
- Alta disponibilidad.

Desventajas:

- Complejidad de gestión.
- Latencia en acceso a datos remotos.

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

Permite que los sistemas manejen crecientes volúmenes de datos y usuarios sin comprometer el rendimiento

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

Facilita la comprensión de datos complejos al transformarlos en gráficos interactivos. Algunos ejemplos:

- Gráficos de dispersión.
- Mapas de calor.

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

Se refiere a las inconsistencias en los datos debido a cambios en las fuentes o formatos, lo que puede complicar el análisis.

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

- En salud para detectar enfermedades
- En ventas para optimizar inventarios o predecir ventas.

18. ¿Que significa el término "valor" en el contexto de Big Data y como se puede obtener valor a partir de los datos?

Es la utilidad obtenida al transformar datos en insights accionables. Por ejemplo, la mejora en la toma de decisiones estratégicas basadas en análisis predictivo.

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

Spark es una plataforma de procesamiento de Big Data en memoria que acelera el análisis de datos. Ventajas:

- Mayor velocidad que Hadoop MapReduce.
- Soporte para múltiples lenguajes (Python, Scala, Java).

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

- Bias algorítmico: Decisiones injustas debido a datos sesgados.
- Consentimiento: Uso de datos sin autorización explícita.
- Transparencia: Falta de claridad en cómo se procesan los datos.