UNIVERSIDAD DE GUADALAJARA

CENTRO UNIVERSITARIO DE CIENCIAS ECONÓMICO ADMINISTRATIVAS

**MAESTRÍA EN CIENCIA DE LOS DATOS**



***Unidad 1. “Trabajo 1 – Fundamentos de Bases de Datos”***

**P R E S E N T A**

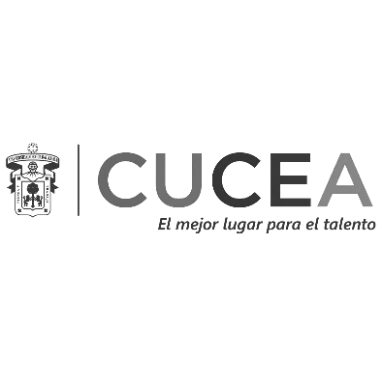
Ilse Arredondo Reyes. Nro Alumno: 323019078.

**Materia: Procesamiento de Grande Bases de Datos**

**Profesor: Ramón Osvaldo Guardado Medina**

**Fecha: 16 de agosto 2023**

**Guadalajara, Jalisco**

Texto

Descripción generada automáticamente con confianza media

Preguntas y respuestas.

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

El término “big data” hace referencia al conjunto de datos de gran volumen y variedad difíciles de procesar y analizar.

Esto también se conoce como “las tres V”.

1. Volumen
2. Variedad
3. Velocidad

Estos conjuntos de datos son tan voluminosos que el software de procesamiento de datos convencional sencillamente no puede gestionarlos. Sin embargo, estos volúmenes masivos de datos pueden utilizarse para abordar problemas empresariales que antes no hubiera sido posible solucionar

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

**Volumen**

Se define como procesar grandes volúmenes de datos no estructurados de baja densidad. Puede tratarse de datos de valor desconocido, como feeds de datos de Twitter, flujos de clics de una página web o aplicación para móviles, o equipo con sensores.

**Velocidad**   
Se define como ritmo al que se reciben los datos y al que se aplica alguna acción. Existen dos tipos:

1. Velocidad de adquisición de datos

2. Velocidad de procesamiento y análisis

La mayor velocidad de los datos normalmente se transmite directamente a la memoria, en vez de escribirse en un disco. Algunos productos inteligentes habilitados para Internet funcionan en tiempo real o prácticamente en tiempo real y requieren una evaluación y actuación en tiempo real.

**Variedad**

La variedad hace referencia a los diversos tipos de datos disponibles.

Existen tres tipos de datos disponibles:

1. Datos estructurados (Bien definidos)
2. Datos no estructurados (Difíciles de manejar y encontrar)
3. Datos semiestructurados (híbridos)

Los tipos de datos no estructurados y semiestructurados, como el texto, audio o video, requieren un preprocesamiento adicional para poder obtener significado y habilitar los metadatos.

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

1. Sensores y Dispositivos IoT: Los dispositivos de Internet de las Cosas (IoT) generan grandes cantidades de datos en tiempo real. Estos dispositivos incluyen sensores en electrodomésticos, vehículos, infraestructuras industriales, dispositivos de seguimiento de salud, etc.
2. Redes Sociales y Medios Digitales:Plataformas como Facebook, Twitter, Instagram y otras redes sociales generan enormes cantidades de datos en forma de publicaciones, comentarios, imágenes y videos compartidos por los usuarios.
3. Transacciones Comerciales y Financieras: Las transacciones en línea, como compras en sitios web y aplicaciones, así como las transacciones financieras en mercados bursátiles y bancos, generan un flujo constante de datos.
4. Registros de Servidores y Dispositivos de Red: Los servidores y dispositivos de red generan registros de actividad, registros de acceso, registros de errores y otros datos relacionados con la administración y el funcionamiento de sistemas informáticos y redes.
5. Datos de Sensores en Ciencia y Experimentación: En campos científicos y de investigación, como la astronomía, la biología y la física, se generan grandes volúmenes de datos a partir de instrumentos de medición y experimentos

**4. Explica la diferencia entre datos estructurados, no estructurados y semi-**

**estructurados en el contexto de Big Data.**

1. Datos estructurados (Bien definidos):

La mayoría de los datos tradicionales son originados por datos de tipo estructurado, datos con formato o esquema fijo, que poseen campos fijos y bien definidos.

1. Datos no estructurados (Difíciles de manejar y encontrar):

Son las estructuras de datos más difíciles de manejar, podemos encontrar entre los datos no estructurados más conocidos: documentos PDF o Word, Audios y videos, Correos electrónicos, Ficheros multimedia de imagen, Artículos y textos, entre otros.

1. Datos semiestructurados (híbridos):

Son un híbrido entre los datos estructurados y los datos no estructurados, ya que no presentan una estructura perfectamente definida pero sí presentan una organización definida en sus metadatos donde describen los objetos y sus relaciones.

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

El procesamiento en tiempo real es un desafío importante para la Big Data debido a la necesidad de analizar y tomar decisiones sobre datos que se generan y fluyen a alta velocidad. Esto implica procesar y analizar datos prácticamente en el mismo momento en que se generan, lo que requiere una infraestructura y tecnología especializada para manejar esta velocidad y volumen de datos.

Ejemplo Plataforma Redes sociales: Una empresa desea realizar un análisis en tiempo real para entender las tendencias y los sentimientos del público hacia un tema específico, debe ser capaz de procesar y analizar todos estos tweets de manera inmediata para la toma de decisiones.

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

Los datos sintéticos son información generada artificialmente que puede utilizarse en lugar de datos históricos reales para entrenar modelos de IA cuando los conjuntos de datos reales carecen de calidad, volumen o variedad. Se utilizan para entrenar modelos de aprendizaje automático, probar software o similares.

Según (Gartner S, 2024) el 60% de los datos utilizados para el desarrollo de soluciones de IA y análisis serán generados sintéticamente, frente al 1% de 2021.

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

La variedad de datos en Big data se describe como las diferentes formas, tipos y fuentes de la información. Esto es importante ya que permite una comprensión más rica y precisa de los fenómenos del mundo real. Sin embargo, esto también puede afectar al análisis y procesamiento de datos debido a que se debe gestionar adecuadamente para obtener lineamientos significativos y tomar decisiones informadas.

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

Apache Hadoop (Infraestructura de datos distribuida) es un framework de software open source o código abierto diseñado para el almacenamiento y procesamiento distribuido de grandes conjuntos de datos en clústeres de servidores. Hadoop se ha convertido en una tecnología fundamental en la Big Data debido a su capacidad para manejar eficazmente el procesamiento y almacenamiento de datos a gran escala, esto permite aumentar la capacidad de procesamiento de manera paralela en clústeres de computadoras actuando como un solo sistema.

La arquitectura de Apache Hadoop está compuesta por varios componentes que trabajan juntos para permitir el almacenamiento y procesamiento distribuido de datos en clústeres de servidores. Los componentes principales de la arquitectura de Hadoop son:

1. Hadoop Distributed File System (HDFS) (Rendimiento lectura y escritura de datos): HDFS es un sistema de archivos distribuido diseñado para almacenar grandes volúmenes de datos de manera eficiente. Divide los archivos en bloques de tamaño fijo y los distribuye en múltiples nodos del clúster. Cada bloque se replica en varios nodos para garantizar la tolerancia a fallos. HDFS es altamente escalable y proporciona un alto rendimiento de lectura y escritura de datos.
2. MapReduce (Procesamiento paralelo de datos): MapReduce es un modelo de programación y un sistema de procesamiento distribuido utilizado para procesar grandes conjuntos de datos. Divide las tareas en dos fases principales: fase de mapeo y fase de reducción. En la fase de mapeo, los datos se procesan en paralelo y se generan pares clave-valor intermedios. En la fase de reducción, los datos se agrupan y se procesan para generar los resultados finales. MapReduce permite el procesamiento paralelo y escalable de datos en clústeres Hadoop.

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

Se define veracidad como al grado de fiabilidad de la información de datos recibida. La veracidad implica que los datos sean precisos y representativos de la realidad que intentan reflejar, sin distorsiones o errores significativos.

La importancia de la veracidad en el contexto de Big data radica en diferentes factores:

1. Para la toma de decisiones precisas (Impacto significativo en los negocios, investigación u otras áreas)
2. Reputación y confianza (Confianza de resultados puede dañar la reputación de la empresa lo que afecta relaciones con clientes y socios)
3. Para un análisis preciso (Interpretación de datos a problemas o tendencias)
4. Eficiencia (Implica Dedicar menos tiempo y recursos en corregir errores o anomalías)
5. Cumplimiento normativo (Implica seguir los lineamientos legales sobre la precisión y confiabilidad de los datos)
6. Identificación de sesgos (Implica analizar datos de manera justa, inclusiva y precisa para la toma de decisiones. Los datos deben ser representativos y precisos para evitar sesgos no deseados en los resultados)

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

Las organizaciones deben implementar medidas de seguridad y privacidad sólidas que incluyan encriptación, autenticación, control de acceso, auditoría, anonimización efectiva y educación para el personal sobre las mejores prácticas de seguridad cibernética.

Algunos desafíos de seguridad y privacidad que se enfrentan en el manejo de Big data son:

1. Fugas de datos y brechas de seguridad (Una fuga de datos o una brecha de seguridad en un conjunto de datos masivo puede resultar en la exposición de información sensible. Ejemplo: En 2017, la seguridad de Equifax expuso datos personales de millones de personas, incluyendo nombres, números de seguridad social y fechas de nacimiento.)
2. Identificación de riesgos y amenazas (Los patrones de ataque pueden ser difíciles de detectar entre el volumen abrumador de datos. Por ejemplo, la detección de un ataque cibernético en un flujo constante de datos de red puede requerir soluciones avanzadas de análisis de comportamiento.)
3. Acceso no autorizado (La gestión de los permisos de acceso a los datos es un desafío en Big Data, especialmente cuando múltiples usuarios y sistemas interactúan con los datos.)
4. Anonimización insuficiente (La combinación de conjuntos de datos aparentemente anónimos puede conducir a la identificación de personas individuales, aunque se puede intentar anonimizar los datos para proteger la privacidad, la reidentificación sigue siendo un riesgo en Big Data. Ejemplo: Un estudio publicado en 2006 por Latanya Sweeney demostró cómo se pudo reidentificar a individuos en una base de datos anónima de registros médicos utilizando información pública.)
5. Riesgos de inferencia: (Los análisis en Big Data pueden revelar información confidencial a través de inferencias. Ejemplo: al analizar patrones de compras y preferencias de usuario, se podría inferir información personal sensible)
6. Protección de la propiedad intelectual (Las empresas que manejan grandes volúmenes de datos deben proteger la propiedad intelectual y los secretos comerciales para evitar cualquier ventaja competitiva y legal.)

**Bibliografía**

1. *Unir, V. (2022, August 4). ¿Qué es el big data y cuál es su importancia? Universidad Virtual. | UNIR Colombia - Maestrías Y Grados Virtuales.* [*https://colombia.unir.net/actualidad-unir/que-es-big-data/*](https://colombia.unir.net/actualidad-unir/que-es-big-data/)
2. *¿Qué es el big data? (n.d.). Oracle Cloud.* [*https://www.oracle.com/mx/big-data/what-is-big-data/*](https://www.oracle.com/mx/big-data/what-is-big-data/)
3. *Coursera. (2023). ¿Qué es big data analytics? Definición, beneficios y más. Coursera.* [*https://www.coursera.org/mx/articles/big-data-analytics*](https://www.coursera.org/mx/articles/big-data-analytics)
4. *Sydle. (2023). Big Data: definición, importancia y tipos.* [*https://www.sydle.com/es/blog/big-data-definicion-importancia-y-tipos-614b791388e600016afa7fc3*](https://www.sydle.com/es/blog/big-data-definicion-importancia-y-tipos-614b791388e600016afa7fc3)
5. *¿Qué son los datos sintéticos? Datos generados para ayudar a tu estrategia de IA. (2022, May 12). IDG Communications S.A.U.* [*https://www.ciospain.es/big-data/que-son-los-datos-sinteticos-datos-generados-para-ayudar-a-tu-estrategia-de-ia*](https://www.ciospain.es/big-data/que-son-los-datos-sinteticos-datos-generados-para-ayudar-a-tu-estrategia-de-ia)
6. *De Ingeniería Del Conocimiento, I. (2021). Seguridad en Big Data, privacidad y protección de datos. Instituto De Ingeniería Del Conocimiento.* [*https://www.iic.uam.es/innovacion/seguridad-big-data/*](https://www.iic.uam.es/innovacion/seguridad-big-data/)