UNIVERSIDAD DE GUADALAJARA

CENTRO UNIVERSITARIO DE CIENCIAS ECONÓMICO ADMINISTRATIVAS

**MAESTRÍA EN CIENCIA DE LOS DATOS**



***Unidad 1. “Trabajo 2 – Fundamentos de Bases de Datos”***

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

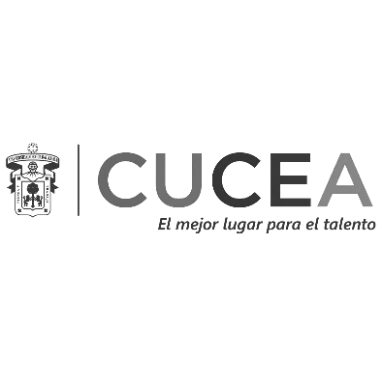
Ilse Arredondo Reyes. Nro. Alumno: 323019078.

**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. 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, son una fuente valiosa de datos en el ámbito del análisis de Big Data debido a la enorme cantidad de información generada por los usuarios en plataformas como Facebook, Twitter, Instagram, LinkedIn y muchas otras. Estas plataformas ofrecen un flujo constante de datos en forma de texto, imágenes, videos, enlaces y más, lo que permite a las organizaciones aprovechar estos datos para realizar análisis significativos y obtener datos relevantes

Según, González, (2019) estos datos, son “rastros”, los cuales son producto de los me gusta, comentarios, fotos, clics, videos, entre otros, producidos por las redes sociales, y son también los registros que se dejan en todas aquellas operaciones tales como transacciones con tarjetas de crédito en los puntos de compra, la utilización de tarjetas en los medios de transportes, las mediciones realizadas por sensores en un mall, datos de teléfonos móviles o los registros efectuados por cámaras de vigilancia, estos rastros en muchos casos se dejan involuntariamente como producto de los quehaceres cotidianos.

Tipos de análisis que se pueden realizar:

1. Análisis de influencia: Identificar a los usuarios más influyentes en una red social específica puede ayudar a las marcas a dirigirse a los defensores clave y establecer asociaciones con ellos.
2. Análisis de campañas de marketing: Las redes sociales son un canal popular para las campañas de marketing.

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

El aprendizaje automático, también conocido como machine learning en inglés, es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender patrones y tomar la mejor decisión a partir de esos datos, sin ser programadas explícitamente para cada tarea. En lugar de utilizar reglas y programación estática, el aprendizaje automático permite a las máquinas mejorar su rendimiento a medida que se exponen a más datos. Cualquier tarea que se base en un grupo de puntos de datos o normas puede automatizarse con el uso del aprendizaje automático, incluso las tareas más complejas como responder llamadas del servicio de atención al cliente y revisar currículums.

Luego, a todo este tipo de datos, aplican técnicas de inteligencia artificial como el machine learning para:

* Encontrar hábitos de consumo y patrones de comportamiento.
* Predecir tendencias.
* Detectar necesidades no satisfechas.
* Recomendar productos de manera personalizada.

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

El almacenamiento distribuido en Big Data se refiere a la práctica de distribuir y almacenar datos en múltiples nodos o servidores, en lugar de centralizarlos en un solo lugar.

*Ventajas del almacenamiento distribuido en Big Data:*

1. Escalabilidad
2. Mayor disponibilidad
3. Mayor rendimiento
4. Menor latencia
5. Diversidad de tecnologías

*Desventajas del almacenamiento distribuido en Big Data:*

1. Complejidad de administración
2. Consistencia y coherencia
3. Requerimientos de red
4. Costos iniciales
5. Seguridad y privacidad

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

La escalabilidad es un aspecto crucial en los sistemas de Big Data debido a la naturaleza inherente de la cantidad de datos que se manejan en estos entornos. Big Data se refiere a conjuntos de datos extremadamente grandes y complejos que exceden la capacidad de las herramientas y enfoques tradicionales para capturar, almacenar, administrar y analizarlos de manera eficiente.

1. Volumen de datos
2. Crecimiento futuro
3. Procesamiento y análisis
4. Respuesta en tiempo real
5. Distribución geográfica
6. Tolerancia a fallos
7. Eficiencia de costos

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

La visualización de datos en Big Data es importante por que nos permite ver datos complejos de una manera mas comprensiva. También nos permite identificar patrones, tendencias y anomalías, además de que sea mas facil de comunicar los datos diferentes audiencias sin limitaciones de hablar el mismo idioma. Los diferentes tipos de visualizaciones son adecuados para diferentes tipos de datos y objetivos de análisis.

Tipos de visualización de datos en el análisis de Big Data:

* Gráficos de barras
* Mapas de calor (Heatmap)

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

La variabilidad se le define como las diferentes interpretaciones que pueden resultar en el proceso, es decir interpretar los datos provenientes de diferentes fuentes.

Esto puede afectar de muchas maneras al análisis, sin embargo, las afecciones principales son: las inconsistencias en los datos y la multitud de dimensiones.

Las inconsistencias de datos se producen cuando existe redundancia de datos, estos se pueden encontrar por métodos de detección de anomalías y valores atípicos para que ocurra cualquier análisis significativo.

Otra afección es la multitud de dimensiones de datos que resultan de múltiples tipos y fuentes de datos dispares. La variabilidad también puede referirse a la velocidad inconsistente a la que se cargan grandes datos en bases de datos.

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

El Big Data, tiene una amplia gama de aplicaciones prácticas en diversas industrias y campos de investigación. El Big data sigue desempeñando un papel crucial en la optimización de procesos, la toma de decisiones informadas y el descubrimiento de conocimientos nuevos y relevantes.

Dos de las muchas aplicaciones prácticas en la industria/investigación que existen son: Salud y Medicina personalizada & Optimización en cadenas de suministros.

1. Salud y Medicina Personalizada:

En la industria de la salud, el Big Data se utiliza para recopilar y analizar datos de pacientes, historiales médicos, resultados de pruebas clínicas, información genética y más. Esto permite obtener una comprensión más profunda de las enfermedades y sus tratamientos, así como identificar patrones y tendencias que podrían no ser evidentes con métodos tradicionales. Un ejemplo es en el 2020 año donde el COVID-19 estaba en su punto más fuerte, el Big data fue aplicado en la salud publica como un modelo predictivo para determinar cuando es probable que se produzca un brote o rebrote entre la población.

1. Optimización de Cadena de Suministro y Logística:

En la industria de la logística y la cadena de suministro, el Big Data se utiliza para recopilar datos en tiempo real sobre el movimiento de bienes, inventarios, transporte y demanda del mercado. El análisis de estos datos permite una gestión más eficiente y optimizada de la cadena de suministro. Un ejemplo sería que las empresas pueden predecir la demanda de productos y ajustar sus niveles de inventario en consecuencia, lo que ayuda a evitar escasez o exceso de stock.

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

En el contexto de Big Data, el término "valor" se refiere a la utilidad, beneficios o conocimientos significativos que se pueden extraer de los datos recopilados y analizados.

Obtener valor a partir de los datos implica realizar análisis inteligentes y significativos para:

1. Toma de Decisiones Informadas (Analizar los datos, las organizaciones pueden tomar decisiones basadas en evidencia en lugar de suposiciones)
2. Identificación de Patrones y Tendencias: Comportamientos de los clientes y otras variables importantes.
3. Personalización y Segmentación: El análisis de datos permite la segmentación de clientes en grupos más específicos según sus preferencias y comportamientos.
4. Predicción y Planificación: Los modelos basados en datos pueden ayudar a prever eventos futuros con cierto grado de precisión.
5. Investigación Científica y Descubrimiento: En la investigación, el análisis de Big Data puede ayudar a descubrir nuevas correlaciones, relaciones y fenómenos en campos como la genómica, la física, la climatología y más.

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

Apache Spark es un sistema de procesamiento de datos en clúster diseñado para el procesamiento y análisis eficiente de grandes volúmenes de datos en tiempo real. Su velocidad de procesamiento es 100 veces mayor que Apache Hadoop si se ejecuta en memoria y 10 veces mayor si se ejecuta en disco, este juega un papel crucial en el procesamiento de Big Data debido a sus características y ventajas que lo diferencian de otras tecnologías:

1. Permite su adaptación a distintas necesidades gracias a que es 100% open source.
2. Simplifica el proceso de desarrollo de soluciones inteligentes.
3. Mejora el desempeño de aplicaciones dependientes de datos.
4. Unifica algoritmos para que trabajen conjuntamente en diversas tareas.
5. Integra dentro de sí el modelado analítico de datos.
6. Otorga escalabilidad en su potencia al introducir más procesadores en el sistema.
7. Reduce los costes al poder utilizarse en hardware estándar de uso común.
8. Promueve workflows basados en Grafos Acíclicos Dirigidos que aceleran el procesamiento.
9. Dispone de una API para Java, Phyton y Scala; también APIs para transformar y manipular datos semiestructurados.
10. Facilita la integración con sistemas de archivos como HDFS de Hadoop, Cassandra, HBase, MongoDB y el S3 de AWS.
11. Ofrece bibliotecas de alto nivel para mejorar la productividad de los desarrolladores.
12. Posee tolerancia a fallos implícita.
13. Combina SQL, streaming y análisis de gran complejidad.

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

El uso de Big Data en la toma de decisiones plantea una serie de desafíos éticos y legales que deben abordarse de manera cuidadosa para garantizar que se respeten los derechos de las personas y se eviten posibles consecuencias negativas.

Existen muchos desafíos tanto éticos y legales relacionados con el uso de Big data, a continuación, se presentan los 3 desafíos más importantes:

1. Privacidad y Protección de Datos:

El uso de Big Data a menudo implica la recopilación, almacenamiento y análisis de una gran cantidad de datos personales. Esto plantea preocupaciones sobre la privacidad y la protección de datos, ya que los individuos pueden sentir que su información personal se está utilizando sin su consentimiento o en formas que podrían ser invasivas.

1. Discriminación y Sesgo:

El análisis de Big Data puede generar resultados sesgados si los conjuntos de datos utilizados reflejan desigualdades existentes o prejuicios históricos. Esto podría llevar a decisiones injustas o discriminatorias.

1. Transparencia y Explicabilidad:

En muchos casos, los modelos de análisis de Big Data pueden ser complejos y difíciles de entender para las personas que son afectadas por las decisiones basadas en esos modelos. Esto plantea problemas de transparencia y explicabilidad. Las personas tienen derecho a comprender cómo se toman las decisiones que las afectan y a obtener explicaciones claras sobre el razonamiento detrás de esas decisiones. La falta de transparencia puede erosionar la confianza en los sistemas basados en Big Data.

Las organizaciones deben implementar prácticas de recopilación y uso de datos éticas, asegurarse de que los algoritmos sean justos y no discriminatorios, y trabajar para mejorar la transparencia y la explicabilidad en las decisiones automatizadas. Además, es esencial estar al tanto de las regulaciones y leyes pertinentes relacionadas con la privacidad y la protección de datos en la región en la que operan.

**Bibliografía**

1. *Universidad de la Rioja AMC Chóez · 2022.* <https://dialnet.unirioja.es/descarga/articulo/8549556.pdf>
2. *Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad. (n.d.).* [*https://www.powerdata.es/big-data*](https://www.powerdata.es/big-data)
3. *Helena. (2020). Aplicaciones principales del Big Data por sectores. Ayuda Ley Protección Datos.* [*https://ayudaleyprotecciondatos.es/big-data/aplicaciones/*](https://ayudaleyprotecciondatos.es/big-data/aplicaciones/)
4. *García Leon, J. (2018). Ética, investigación, bases de datos. Javier García León. Recuperado de* [*https://javiergarcialeon.net/2018/11/12/etica-investigacion- bases-de-datos/*](https://javiergarcialeon.net/2018/11/12/etica-investigacion-%20bases-de-datos/)*.*
5. *¿Qué es el aprendizaje automático? | Glosario*. (s. f.). HPE LAMERICA <https://www.hpe.com/lamerica/es/what-is/machine-learning.html>