Posgrados
CUCEA
El mejor lugar para el talento

Procesamiento de Grandes Bases de Datos

Actividad 1

Profesor: Dr. Osvaldo Guardado

Alumno: Oscar Fernando Flores Garcia
Código: 323019043

Zapopan, Jalisco a 23 de Agosto del 2023

1. Define qué es big data y menciona al menos tres características clave.

Big data se refiere a los conjuntos de datos de volúmenes masivos que pueden contener información estructurada, semiestructurada y no estructurada; son tan grandes que no pueden ser manejados por el software de procesamiento de datos tradicional.

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

Volumen: Grandes volúmenes de datos que pueden ser de baja densidad y datos no estructurados.
Variedad: Se refiere a los diferentes tipos de datos disponibles, incluyendo estructurados como los formados en bases de datos SQL, así como no estructurados o semi estructurados como video, audio, documentos, entre otros.
Velocidad: En el big data los datos se generan a una velocidad superior a aplicaciones tradicionales.

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

    1. Información de una red social.
    2. Logs generados por múltiples sistemas.
    3. Información de tráfico en tiempo real.
    4. Información de monitoreo de clima en campos agrícolas.
    5. Monitoreo de APIs.

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

Tipo de dato	Estructurado	Semiestructurado	No estructurado
Concepto	Es información organizada para un fácil análisis, contiene sus filas y columnas para cada dato.	Información que no reside en una base de datos relacional pero aun así contiene alguna estructura donde se pueden extraer los elementos del dato.	Datos que no están organizados de una manera o modelo predefinido.
Ejemplos	- Bases de datos SQL - CSV	- XML - HTML	- Música - Video - Documentos

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

El desafío del procesamiento de datos en tiempo real en un entorno de Big Data reside en la ingesta, el procesamiento y almacenamiento de datos. El procesamiento se debe de realizar

de manera que no bloquee la ingesta de datos, el almacenamiento debe admitir la escritura de grandes volúmenes de datos.

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

Son modelos generados a partir de datos generados artificialmente para entrenar modelos en datos sintéticos en lugar de o como complemento a datos históricos reales.

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

El contar con datos variados en el big data nos ofrece una perspectiva más amplia y puede ayudar a crear sesgos en modelos predictivos. Entre mas variada sea la información podemos realizar análisis multivariados más complejos.

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

Hadoop es un framework de código abierto que permite usar modelos sencillos de programación para almacenar y procesar de forma distribuida grandes conjuntos de datos de distintos clusters de ordenadores.

HDFS: Es el componente principal del ecosistema de Hadoop. Es el sistema de archivos distribuidos que franquean el acceso de alto rendimiento a los datos de las aplicaciones sin tener que definir esquemas de antelación.

MapReduce: Este modelo de programación permite procesar los datos a gran escala y de manera paralela con un algoritmo distribuido.

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

Se refiere al sesgo, ruido y alteración de los datos. Su importancia radica en que se deben seleccionar los datos de manera consciente al problema que se esté tratando y en medida de la confianza que se tenga en los mismos.

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

Existen múltiples retos en diferentes etapas, obtención, procesamiento y almacenamiento. Algunos ejemplos son:

- donde obtenemos la información y que nivel de detalle es necesario para nuestro problema?
- En el momento del procesamiento, que tanto se expone a servicios de terceros y propios?
- El almacenamiento de datos puede estar sujeto a regulaciones, por ejemplo, los datos bancarios de clientes europeos no deben residir fuera de la unión europea.

Referencias

Amazon Web Services. (s.f.). *Hadoop y Spark: diferencia entre los marcos de Apache*. AWS. Consultado el 23 de Agosto, 2023, de https://aws.amazon.com/es/compare/the-difference-between-hadoop-vs-spark/

Geeksforgeeks. (2023, March 6). *Difference between Structured, Semi-structured and Unstructured data*. GeeksforGeeks. Consultado el 18 de Agosto ,2023, de https://www.geeksforgeeks.org/difference-between-structured-semi-structured-and-unstructured-data/

GeeksforGeeks. (2023, May 4). *Advantages of Distributed database*. GeeksforGeeks. Consultado el 21 Agosto, 2023, de https://www.geeksforgeeks.org/advantages-of-distributed-database/

Google. (s.f.). *¿Qué es Hadoop?* Google Cloud. Consultado el 21 Agosto, 2023, de https://cloud.google.com/learn/what-is-hadoop?hl=es

Google. (s.f.). *What is Apache Spark?* Google Cloud. Consultado el 23 Agosto, 2023, de https://cloud.google.com/learn/what-is-apache-spark

IBM. (s.f.). *¿Qué es Machine Learning? - México*. IBM. Consultado el 21 Agosto, 2023, de https://www.ibm.com/mx-es/analytics/machine-learning

Oracle. (s.f.). *What Is Big Data?* Oracle. Consultado el 18 Agosto, 2023, de https://www.oracle.com/big-data/what-is-big-data/