

Lectura #1

Bases de Datos II

Luis Diego Delgado Muñoz

Prof. Nereo Campos

Fecha de Entrega: 23/02/2024

Responda las siguientes preguntas:

1. ¿En qué consisten datos estructurados, semiestructurados y no estructurados?
Comente ejemplos de estos tipos de datos. **R/** Primeramente, los datos estructurados, están altamente organizados y son fácilmente comprensibles por algoritmos de machine learning. Los datos estructurados se utilizan para lo que comunmente conocemos como bases de datos relacionales y SQL es el lenguaje de programación utilizado para gestionar datos estructurados. Algunos ejemplos de datos estructurados incluyen fechas, nombres, direcciones, números de tarjetas de crédito, entre otros. En medio de los estructurados y no estructurados tenemos los semi-estructurados, estos son la transición entre los datos estructurados y no estructurados. No tienen un modelo de datos predefinido y son más complejos que los datos estructurados, pero más fáciles de almacenar que los datos no estructurados. Algunos ejemplos son archivos JSON, CSV, XML, entre otros. Por último, los datos no estructurados, no pueden ser procesados y analizados mediante herramientas y métodos de datos convencionales. Dado que los datos no estructurados no tienen un modelo de datos predefinido es mejor gestionarlos en bases de datos no relacionales (NoSQL). Algunos ejemplos son archivos multimedia, datos de sensores de IOT, entre otros. (IBM, 2021)
2. ¿En qué consisten datos de series de tiempo? ¿Se consideran logs, datos de series de tiempo? **R/** Según IBM, los datos de series de tiempo consisten en una colección ordenada de medidas tomadas a intervalos regulares, por ejemplo, precios diarios de acciones o datos de ventas semanales (IBM, 2021). Los datos que se capturan pueden ser de cualquier tipo, según lo que necesite capturar. Con respecto a los logs, sí se considerarían datos de series de tiempo ya que son datos que se capturan en ciertos intervalos de tiempo regulares para poder analizar el comportamiento de el software, según los datos que se toman en el tiempo definido.
3. ¿Comente diferencias entre Lake house, Data warehouse y Data mart? **R/** Estos tres patrones de arquitectura son muy similares, sin embargo, la principal diferencia es el uso y la funcionalidad. Un Data Mart se utiliza principalmente para trabajar solo con datos para un área en particular, así que realizar queries entre diferentes áreas suele ser complejo. A diferencia de los data marts, los data warehouses son capaces de analizar y recibir gran cantidad de datos diferentes para realizar análisis y descubrimiento de patrones en los datos recopilados. Para terminar, un lake house combina los mejores elementos de los data warehouses y los data lakes. Este permite mover datos entre los warehouses y lakes de una manera más eficiente y veloz.
4. ¿En qué consiste Row-oriented Column-oriented databases? Suponiendo que existe una tabla en una base de datos relacional con 10 columnas cuyos nombres son column1, column2, ..., column10, ¿Una consulta como "SELECT column1, column2 FROM

tabla" se vería mas beneficiada por Row-oriented o Column-oriented? Explique.

R/ Las **row-oriented** DB son bases de datos que se almacenan en bloques completos todas las filas, a diferencia de los **column-oriented** DB que almacenan las columnas en cada en un bloque físico por cada columna. Tomando esto en cuenta, para mí se vería más beneficiado usando una modalidad column-oriented, ya que si fuera row, tendría que visitar todas las columnas aunque solo se quisieran consultar algunas nada más.

Referencias Bibliográficas

- IBM, E. (2021, Junio 29). Structured vs. unstructured data: What's the difference? IBM Blog. <https://www.ibm.com/blog/structured-vs-unstructured-data/>
- IBM, E. (2021, Agosto 17). Time series data. Time Series Data. <https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-time-series-data>