Lectura #1 Solución

Kevin Córdoba Chevez - 2020100920

1. ¿En qué consisten datos estructurados, semiestruutados y no estructurados? Comente ejemplos.

1.1 datos estructurados.

"Los datos estructurados se clasifican con mayor frecuencia como datos cuantitativos, y es el tipo de datos con el que la mayoría de nosotros estamos acostumbrados a trabajar"(Juan, 2020), como se puede apreciar, los datos estructurados son datos cuantitativos, estos son los datos que se utilizan en el día a día y se puede pensar en ellos como los datos que se pueden almacenar en una base de datos relacional.

Un ejemplo de estos datos puede ser: nombres y fechas.

1.2 datos no estructurados:

"Los datos no estructurados se clasifican con mayor frecuencia como datos cualitativos y no pueden procesarse y analizarse utilizando herramientas y métodos convencionales."(Juan, 2020), se puede ver que los datos no estructurados son datos cualitativos y estos al contrario de los datos estructurados no se almacenan en una base de datos relacional, sino que una no relacional es lo más adecuado.

Un ejemplo de estos datos: audio y video.

1.3 datos semiestructurados:

"información que no reside en una base de datos relacional ni en ninguna otra tabla de datos, pero que, sin embargo, tiene algunas propiedades organizativas para facilitar su análisis, como las etiquetas semánticas."(Juan, 2020), se puede apreciar que a diferencia de los anteriores tipos de datos, estos semiestructurados no son los que se almacenan en una base de datos, ya sea relacional o no, se encuentran en medio de ambos y poseen características de cada uno.

Un ejemplo de estos datos sería el cógigo HTML.

2. ¿En qué consisten datos de series de tiempo? ¿Se consideran logs datos de series de tiempo?

"Por serie de tiempo nos referimos a datos estadísticos que se recopilan, observan o registran en intervalos de tiempo regulares (diario, semanal, semestral, anual, entre otros)."(UNISON, s.f.), se puede observar que datos de series de tiempo se refiere a los datos que se recopilan sobre el tiempo, los cuales permiten realizar estadísticas y comparaciones.

Según la definición anterior considero que los log son efectivamente series de tiempo, ya que su función consiste en recopilar y almacenar información de un programa o base de datos en el tiempo.

3. Comente diferencias entre Lake house, Data warehouse y data mart

Según la lectura realizada en la plataforma Perusall,

-Un lake house es un patrón arquitectónico que combina los mejores delementos de data warehouses y data lakes, además permiten consultar datos a través del data warehouse. Permiten ganar conocimientos más rápido y a mayor profundidad que no sería posible d otra manera.

Con esta arquitectura se puede almacenar datos en formatos open file en su data lake y consultarlos en el lugar mientras se unen con datos del data warehouse.

-Un data warehouse permite correr análisis veloces en grandes volúmens de datos y descubrir patrones ocultos en los datos usando herramientas de BI. Los científicos de datos utilizan un data warehouse para ejecutar análisis offline y detectar tendencias.

-Un data mart es una forma simple de data warehouse, enfocado en un área funcional específica o en un tema de importancia. Sin embargo, debido a que están enfocados en áreas funcionales específicas, se pueden volver díficiles de distribuir.

4. ¿En qué consiste Row-oriented Column-oriented databases?

-Row-oriented databases:

Usualmente, almacenan filas completas en un bloque físico. El alto rendimiento para operaciones de lectura se consigue mediante índices secundarios.

Se usan tradicionalmente para data warehousing, pero encajan mejor en procesamiento transaccional (OLTP) que para análiticas.

-Colum-oriented databases:

Organizan cada columna en su propio set de bloques físicos, en lugar de empaquetar todas las filas en un solo bloque. Esto les permite ser más eficiente con inputs/outputs (I/O) para consultas de solo lectura, debido a que solo tienen que leer esas columnas accesadas por una consulta del disco. Este acercamiento hace que las bases column-oriented sean una mejor opción para data warehousing que las bases row-oriented.

4.1 Suponiendo que exista una tabla en una basse de datos relacional con 10 columnas cuyos nombres son column1, column2, ..., column10, ¿una consulta como "SELECT column1, column2 FROM tabla", se vería más beneficiada por Row.oriented o Column-oriented? Explique.

Según lo comprendido en la lectura yo considero que para este caso es mucho mejor seleccionar la opción de Column-oriented, ya que al estar separada cada columna individualmente consultarlas resulta mucho más veloz que si estuvieran todas en un solo bloque como en el caso de las row-oriented. Además, esta es una consulta de solo lectura, y es en estas donde más brilla la opción de column-oriented, aparte de de tener una mejor compresión de los datos.

Referencias

Juan.(2020). Diferencia entre datos estructurados y no estructurados. Ayuda Ley Protección Datos. Recuperado de:

https://ayudaleyprotecciondatos.es/bases-de-datos/diferencias-entre-datos-estructurados-y-no-estructurados/#:~:text=Los%20datos%20estructurados%20son%20modelos,y%20an%C3%A1lisis%20es%20m%C3%A1s%20dif%C3%ADcil.

UNISON.(s.f.). Series de Tiempo. Recuperado de:

http://www.estadistica.mat.uson.mx/Material/seriesdetiempo.pdf