Quiz #1

Fecha: 26/08/2022
Autor: Deyan Sanabria Fallas #2021046131

1. Explique en que consisten los siguientes conceptos

Data Warehouse

Los Data Warehouse son almacenes de datos especializados para el análisis de datos, se caracterizan por ser lentos para Queries SQL y comunmente necesitan de un intermediario.

Data Lake

A diferencia del Data Warehouse, los data lakes son almacenes de datos raw masivos preparados con Multipart upload y Multipart Download.

Data Mart

El Data Mart es básicamente un Data Warehouse centrado a un tema o área en específico. Se pueden tener varios de estos para dividir en secciones de una empresa.

2. ¿De que forma se benefician las aplicaciones del uso de Columnar Storage?

El columnar Storage, a diferencia del Row Storage, almacenan los datos por columna. Es más, fácil de comprender entendiendo las debilidades del Row Storage. El Row Storage tiene la desventaja de tener que tratar con diferentes tipos de datos por columna, lo que no permite la flexibilidad y compresión que si permite el Columnar Storage, un claro ejemplo sería una columna con varios tipos diferentes donde se encuentra un varchar, en este caso, si se usa un varchar de 512 y se usan solo 50 caracteres en promedio, se está perdiendo mucho espacio y este espacio se suele dejar fijo aunque no se use porque así se sabe dónde inicia y termina ese tipo de dato para esa fila. Con el Columnar Storage, se guarda por columnas por lo que todos los datos que hay dentro son del mismo tipo entonces no existe esta desventaja y es más fácil para la compresión, además que, si se necesita buscar datos por columna que es lo que generalmente pasa, esto se puede hacer de forma muy rápida.

3. ¿En que consiste streaming y batch processing?

Streaming Processing

El Streaming Processing es simplemente un flujo de datos continuo, un gran ejemplo, pueden ser servicios de streaming como youtube, donde se necesita un paso de datos continuo para visualizar videos. Por ende, el proceso de datos se hace de manera continua a diferencia del siguiente tipo.

Batch Processing

El Batch Processing, a diferencia del Streaming Processing el cual es un flujo continuo, se tratan de datos que se van procesando por grupos, de ahí el nombre “batch” que en español significaría algo como “lotes” o “grupos”.

4. ¿En que consiste datos estructurados, semi estructurados y no estructurados?

Datos Estructurados

Los datos estructurados son aquellos que tienen un patrón definido, estos se pueden tabular como en una base de datos o una hoja de Excel y siempre contienen los mismos tipos de datos (columnas).

Datos No Estructurados

Los datos no estructurados no tienen ningún patrón definido, estos no suelen tener forma de siquiera darles un formato tabular y no se pueden asociar de alguna manera.

Datos Semi-estructurados

Los datos semi-estructurados, son una mezcla de los tipos anteriores, tienen ciertos patrones para relacionarlos de alguna forma, pero siguen teniendo datos sin ninguna estructura a los que se pueda asociar.

Referencias

Amazon Web Services. (15 de Enero de 2021). *Data Warehousing on AWS*. Obtenido de Amazon Web Services: https://docs.aws.amazon.com/whitepapers/latest/data-warehousing-on-aws/data-warehousing-on-aws.html