

1. Explique en que consisten los siguientes conceptos:

a. **Data Warehouse:** Es un almacén de datos especializado, particularmente para generar *insights*. Estos consisten en conclusiones a las que se llega a partir del análisis de datos. Los data warehouses se encuentran optimizados para operaciones de escritura por lotes y una alta lectura de volúmenes de datos.

b. **Data Lake:** Contienen "raw data", es decir, datos sin procesar. Estos son almacenados en un object storage en cantidades masivas y a un bajo costo.

c. **Data Mart:** Son una versión simple de almacén de datos centrada en un área funcional específica o tema. Los data marts son fáciles de diseñar, construir y administrar.

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

Las bases de datos orientadas a columnas organizan cada columna en su propio conjunto de bloques físicos en lugar de empaquetar todas las filas en un bloque. Esto beneficia las aplicaciones ya que les permite ser más eficientes para consultas de solo lectura. Además, otro gran beneficio es una compresión mejorada. Como cada columna está empaquetada en su propio conjunto de bloques, cada bloque físico contiene el mismo tipo de datos. Cuando esto ocurre y toda la información es del mismo tipo, es posible usar algoritmos de compresión extremadamente eficientes, de manera que se necesita menos almacenamiento comparado a una base de datos que usa Row Storage. [1]

3. ¿En que consiste streaming y batch processing?

Streaming: Los datos de streaming aparecen en un flujo constante, no se tiene un principio ni final propiamente para los datos y no es posible definir su tamaño en disco. Por esto, el procesamiento en streaming consiste en recolectar, almacenar y procesar esta información continuamente.

Batch Processing: Consiste en procesar información por lotes, posiblemente en intervalos definidos.

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

Datos estructurados: Estos datos siguen un formato de esquema tabular, es decir, columnas con sus respectivos tipos de datos.

No estructurados: Estos datos no siguen ningún patrón. Algunos ejemplos pueden ser texto sin estructura, audio y/o video.

Semi estructurados: Estos datos no necesariamente tienen los mismos atributos.

Bibliografía:

1. Amazon Web Services. (2021, Enero). Data Warehousing on AWS. Recuperado el 26 de Agosto, 2022