Prueba Corta # 2

Tecnológico de Costa Rica
Escuela de Ingeniería en Computación
Bases de datos II (IC 4302)
Primer Semestre 2023
Fecha de entrega: **14/03/23 antes de las 11:59 pm**
Forma de entrega: **Email al profesor siguiendo los lineamientos del programa de curso, adjuntando documento y link al repositorio.**
Formato: **Markdown**
Nombre Archivo: **pc2.md**

Gerald Núñez Chavarría – 2021023226

1. **Explique en qué consisten los siguientes conceptos:**
    a. Data Warehouse
    Un Data Warehouse es un depósito centralizado de datos procedentes de múltiples fuentes, diseñado para dar soporte a actividades de BI como la elaboración de informes, el análisis y la minería. Contienen datos históricos, estructurados para una consulta y un análisis eficaces, con el fin de respaldar los procesos de toma de decisiones. Son utilizados por organizaciones con grandes cantidades de datos en múltiples sistemas para integrar y poner los datos a disposición de los usuarios empresariales.

    b. Data Lake
    El Data Lake es un repositorio para almacenar datos sin procesar (raw data) en su formato nativo para soportar el procesamiento de big data, el aprendizaje automático y las cargas de trabajo de análisis avanzado. A diferencia del Data Warehouse, este almacena los datos en su forma original, lo que proporciona flexibilidad y escalabilidad para procesar grandes volúmenes de datos de varias fuentes. Data Lake puede utilizarse para respaldar la analítica avanzada, el aprendizaje automático y el análisis exploratorio, donde los científicos y analistas de datos pueden descubrir nuevos patrones y perspectivas.

    c. Data Mart
    Un Data Mart es un subconjunto de un Data Warehouse que sirve a una unidad de negocio o departamento específico dentro de una organización. Contiene un subconjunto de datos más pequeño que un Data Warehouse completo y está diseñado para dar soporte a requisitos empresariales específicos, como la elaboración de informes o el análisis. Los Data Marts pueden crearse para funciones específicas e integrarse con otros Data Marts o con un Data Warehouse más grande. Al igual que los Data Warehouse, los mercados de datos admiten diversas actividades de BI.

2. **¿De qué forma se benefician las aplicaciones del uso de Columnar Storage?**
    Las aplicaciones se pueden beneficiar del Columnar Storage de varias formas, por ejemplo, los sistemas de Columnar Storage están diseñados para reducir el tiempo necesario para leer datos específicos de una tabla, ya que solo se accede a las columnas relevantes en lugar de leer toda la fila, lo que reduce la carga de lectura de E/S. También, los datos almacenados se pueden comprimir más eficazmente que en un sistema Row Storage, lo que reduce la cantidad de espacio de almacenamiento necesario. Además, la arquitectura de almacenamiento columnar permite una mayor paralelización de las operaciones de consulta, lo que puede acelerar significativamente el tiempo de respuesta para consultas complejas. En general, el uso de Columnar Storage permite una mayor eficiencia y escalabilidad en la manipulación y el análisis de grandes conjuntos de datos.

3. **¿En qué consiste streaming y batch processing?**
   El streaming es un enfoque de procesamiento de datos en tiempo real, en el que los datos se procesan a medida que se generan. El flujo de datos se procesa en pequeñas porciones, llamadas micro lotes, en lugar de procesar grandes volúmenes de datos en un solo lote. Los datos se procesan continuamente a medida que se reciben, lo que permite una toma de decisiones en tiempo real. Aunque procesar datos en tiempo real suena fabuloso, la principal desventaja del streaming es que suele ser muy costoso a nivel de recursos, sobre todo al tener volúmenes grandes de datos que procesar.

   El procesamiento en batch implica la recopilación y el procesamiento de grandes volúmenes de datos en un solo lote. Los datos se recopilan durante un período de tiempo determinado y se procesan en un lote al final de ese período. Esto se hace típicamente en horarios programados e intervalos regulares, para procesar grandes volúmenes de datos.

4. **¿En qué consiste datos estructurados, semi estructurados y no estructurados?**
   Los datos estructurados se organizan en una estructura predefinida, como una tabla de una base de datos relacional, y tienen un formato consistente y bien definido. Se pueden consultar fácilmente y analizar con herramientas de análisis de datos tradicionales, como consultas SQL.

   Los datos semi estructurados no tienen un formato predefinido y no se ajustan a las estructuras tradicionales de bases de datos relacionales. Los datos semi estructurados a menudo se representan en formatos como JSON o XML y pueden tener una estructura jerárquica.

   Los datos no estructurados no tienen una estructura predefinida y no se pueden almacenar en una base de datos relacional. Ejemplos de datos no estructurados incluyen texto, imágenes, audio y video. Estos datos son difíciles de analizar y procesar con herramientas de análisis de datos tradicionales. Las técnicas de procesamiento de lenguaje natural, el aprendizaje automático y el análisis de imágenes y video se utilizan para extraer información de los datos no estructurados.