TEC | Tecnológico de Costa Rica

Instituto Tecnológico de Costa Rica.

Campus Tecnológico Central Cartago

Escuela de Ingeniería en Computación.

Curso: Bases de Datos II IC-4302

I Semestre 2023.

Prueba Corta 2

Fecha de entrega: 14/03/2023

Profesor:

Nereo Campos Araya.

Estudiante:

Diana Sanabria Calvo 2021436548

Índice

Bases de Datos II
Prueba Corta 2

Elaborado por Diana Sanabria Calvo.

1. Instrucciones

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**

---

1. Explique en que consisten los siguientes conceptos:
    a. Data Warehouse
    b. Data Lake
    c. Data Mart
2. ¿De que forma se benefician las aplicaciones del uso de Columnar Storage? Explique.
3. ¿En que consiste streaming y batch processing?
4. ¿En que consiste datos estructurados, semi estructurados y no estructurados?

Figura 1: *Instrucciones Generales*

2. Explique en que consisten los siguientes conceptos:

2.1. Data Warehouse

Permite la integración, consolidación y análisis de cantidades significativas de datos de varias fuentes para producir información valiosa y práctica para la toma de decisiones.También podemos decir que

es un almacén de datos, un tipo de base de datos creada específicamente para contener grandes cantidades de datos de una organización y gracias a esto permitir el análisis de esos datos para producir información que sea relevante y útil.

2.2. Data Lake

Es un repositorio centralizado y escalable que permite el almacenamiento de cantidades considerables de datos estructurados, semiestructurados y no estructurados de diversas fuentes en sus formatos originales. Un Data Lake a diferencia de un Data Warehouse, permite a los usuarios explorar los datos de manera más libre y flexible porque no requiere una estructuración previa de los datos y puede almacenar datos sin procesar.

2.3. Data Mart

Un Data Mart es una parte de un Data Warehouse que incluye datos particulares para un equipo o departamento y está destinado a ser utilizado por un conjunto particular de usuarios para analizar y tomar decisiones. Un Data Mart se crea para ser utilizado por un grupo más pequeño de usuarios y para un propósito específico, en contraste con un Data Warehouse que puede incluir datos de toda la organización.

Se puede desarrollar un Data Mart en AWS utilizando servicios como Amazon Redshift, que le permite crear subconjuntos de un Data Warehouse y asignar permisos de acceso a usuarios específicos en función de sus necesidades y roles en la organización. Como resultado, la gestión de datos es más efectiva y las necesidades únicas de cada departamento de la organización están mejor respaldadas.

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

Mayor rendimiento de consultas: debido a que los datos se organizan por columnas en lugar de filas, las bases de datos que utilizan almacenamiento en columnas pueden procesar consultas de manera más rápida y efectiva. Como resultado, los datos se pueden comprimir de manera más efectiva y se pueden leer más rápidamente.

Ahorro en almacenamiento: las bases de datos que usan almacenamiento en columnas pueden almacenar datos de manera más compacta porque los valores repetidos en una columna solo se almacenan

una vez. En consecuencia, se puede ahorrar espacio de almacenamiento y se pueden reducir los costos relacionados con el mismo.

Compatibilidad con análisis avanzados: las bases de datos de almacenamiento en columnas se adaptan perfectamente para proporcionar operaciones complejas de análisis de datos, procesamiento de datos en paralelo, filtrado de datos y agregación.

4. ¿En que consiste streaming y batch processing?

Streaming es una técnica de procesamiento de datos en tiempo real, donde los datos se procesan y se analizan a medida que se generan, lo que permite una respuesta rápida a los cambios en los datos. En un sistema de streaming, los datos se transmiten continuamente desde la fuente de origen y se procesan en tiempo real utilizando herramientas de procesamiento de flujo, como Apache Kafka, Apache Flink o Amazon Kinesis.

Batch processing es una técnica de procesamiento de datos que implica la recopilación de grandes cantidades de datos y su procesamiento en un solo batch en lugar de en tiempo real. En un sistema de batch processing, los datos se recolectan en un almacenamiento de datos centralizado, como un Data Warehouse o un Data Lake y luego se procesan en grandes lotes utilizando herramientas de procesamiento de batch, como Apache Hadoop, Apache Spark o AWS Batch.

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

Los datos estructurados como bien dice su nombre, se mantienen en un formato estructurado, como en una base de datos relacional. Los datos estructurados se guardan en tablas con columnas y filas que tienen un esquema predeterminado fijo. Debido a esto, los lenguajes de consulta como SQL simplifican la organización y consulta de los datos.

Los datos que carecen de un formato estructurado establecido pero que aún tienen una estructura reconocible se denominan datos semiestructurados. Los datos semiestructurados se almacenan en formatos como JSON o XML, que proporcionan una estructura jerárquica para los datos. Aunque los datos no tienen un esquema establecido, las herramientas de procesamiento de datos semiestructurados como Apache Hive aún pueden organizarlos y consultarlos. Los documentos HTML, las fuentes RSS y los registros de eventos son algunos ejemplos de datos semiestructurados.

Datos no estructurados: son datos que no tienen una estructura fija o reconocible. Los datos no

estructurados se almacenan en formatos como texto, audio o video y no se pueden organizar o acceder de manera efectiva mediante herramientas para procesar datos estructurados o semiestructurados. Para encontrar patrones y extraer información pertinente de datos no estructurados, se pueden utilizar técnicas de aprendizaje automático o herramientas de procesamiento de lenguaje natural como Amazon Comprehend. El correo electrónico, las redes sociales, las imágenes y los videos son algunos tipos de datos no estructurados.