OMNICLOUD Prueba técnica v2

Perfil: DBA mid

Prueba Técnica: Docker, ETL y Big Data

Objetivo

Crear un entorno que incluya:

1. Un contenedor con PostgreSQL.
2. Un contenedor con MongoDB.
3. Un proceso ETL que integre ambas bases de datos.
4. Una base de datos de Big Data configurada para alta disponibilidad.
5. Una base de datos inicial en PostgreSQL, que ya se proporcione al candidato para que sea optimizada.

Instrucciones

PostgreSQL

1.- Normaliza e indexa la siguiente base de datos:

```
CREATE DATABASE db;

CREATE TABLE orders (
   order_id SERIAL,
   customer_name TEXT,
   product_name TEXT,
   product_price NUMERIC,
   order_date TIMESTAMP
);

INSERT INTO orders (customer_name, product_name, product_price, order_date)
VALUES
   ('John Doe', 'Laptop', 1200.50, '2024-01-01 10:00:00'),
   ('Jane Smith', 'Smartphone', 800.75, '2024-01-02 11:30:00'),
   ('Alice Johnson', 'Tablet', 300.00, '2024-01-03 15:45:00'),
   ('Bob Brown', 'Monitor', 150.99, '2024-01-04 09:20:00'),
   ('Carol White', 'Keyboard', 50.00, '2024-01-05 14:10:00');
```

2.- Crea un contenedor de docker, donde se guarde la base de datos normalizada.

MongoDB

1. Configura un contenedor con MongoDB.
2. Crea una colección inicial llamada logs que contenga registros aleatorios como datos JSON.
3. Genera una estructura para cargar los datos transformados desde PostgreSQL como parte del proceso ETL.

ETL

1. Crea un script que:
    - Extraiga datos de la tabla users en PostgreSQL.
    - Transforme los datos aplicando reglas como:
        - Cambiar el formato de los nombres a mayúsculas.
        - Modificar los correos para usar el dominio example.com
        - Agregar un código único por usuario.
    - Cargue los datos transformados en una nueva colección llamada users_transformed en MongoDB.
2. Configura una tarea para que el proceso ETL se ejecute automáticamente después de los respaldos.

Big Data (opcional, es un plus)

1. Configura un contenedor para Apache Kafka o Elasticsearch:
    - Apache Kafka:
        - Configurarlo para recibir datos de eventos, como logs o registros de usuarios.
        - Proporciona un productor básico para enviar eventos al tópico user_events.
    - Elasticsearch:
        - Configurarlo para indexar y analizar datos provenientes de los eventos.

Respaldo

1. Configura un script para respaldos automáticos de PostgreSQL y MongoDB cada hora.
2. Los respaldos deben guardarse en una carpeta compartida entre los contenedores y el host.

Consultas (PostgreSQL)

Crea las siguientes consultas

1. Lista de todos los pedidos junto con nombre del cliente y el producto en un rango de fechas.
2. Calcula el total de ventas por cliente.
3. Encuentra a los 3 mejores clientes, que tengan el mayor gasto.
4. Verifica que los índices de tus tablas estén siendo utilizados.
5. Muestra los nombres de los clientes que no han realizado ningún pedido.

Entrega

1. Un archivo docker-compose.yml que configure todos los contenedores:
    - PostgreSQL.
    - MongoDB.
    - Kafka o Elasticsearch.
    - ETL.
2. Scripts necesarios:
    - Respaldo automático.
    - Normalización de la base de datos.
    - Proceso ETL.
3. Documentación clara que incluya:
    - Instrucciones para ejecutar el entorno.
    - Detalles sobre cómo realizar la normalización.
    - Descripción del proceso ETL.
    - Uso de Kafka o Elasticsearch.
4. Subir el resultado a un repositorio público de Github
5. Enviar por correo electrónico a la persona con quien estás llevando el proceso el enlace al repositorio del proyecto.

Criterios de Evaluación

1. PostgreSQL:
    - Correcta configuración y creación de la base de datos inicial.
    - Habilidad para normalizar y optimizar una base de datos.
    - Configuración adecuada para alta disponibilidad (por ejemplo, replicación).
2. MongoDB:
    - Configuración funcional y uso adecuado de colecciones.
    - Carga exitosa de datos transformados desde el proceso ETL.
3. ETL:
    - Correcta extracción, transformación y carga de datos.
    - Automatización del proceso ETL.
4. Big Data:
    - Configuración funcional de Kafka o Elasticsearch.
    - Manejo correcto de datos en tiempo real.
5. Respaldo y Automatización:
    - Implementación funcional de respaldos automáticos.
    - Uso correcto de volúmenes para compartir datos entre contenedores.
6. Documentación:
    - Claridad en las instrucciones.
    - Facilidad para reproducir el entorno en otra máquina.