Adriel Araya Vargas 2019312845

IC-4302 Bases de Datos 2

Examen

I Semestre

10 de junio de 2023

Pregunta 1 (60 pts)

Aproximadamente para el año 23651 de nuestra era y durante el apogeo del imperio galáctico, el matemático Hari Seldon ha desarrollado su teoría llamada Psicohistoria, mediante esta, ha podido predecir con un grado de confianza bastante alto la caída de la civilización seguida de un periodo de barbarie, con el fin de reducir este periodo de barbarie, este ha desarrollado un plan y como parte de este, se encuentra la conformación de la Enciclopedia Galáctica, la cual de acuerdo con el divulgador científico Carl Sagan es un sugerente proyecto del saber colectivo de las civilizaciones avanzadas del universo. Usted ha sido escogido como líder técnico del equipo que se encargará de implementar la base de datos que mantendrá esta información con alta disponibilidad y con un mecanismo adecuado para navegar los datos y realizar búsquedas. Es importante mencionar:

• La tecnología en bases de datos SQL y NoSQL no han cambiado desde el año 2023.

• No existe restricción en cuanto a dinero que se puede invertir en el proyecto.

• Los proveedores de Cloud siguen existiendo y ahora han expandido sus ubicaciones en prácticamente todo el universo conocido.

• Los productos ofrecidos en los proveedores de Cloud para el 2023 siguen siendo ofrecidos para el año 23651.

• Se tiene que permitir full text search sobre la información en la Enciclopedia Galáctica.

• Se tienen que establecer relaciones entre los diferentes elementos de información de forma tal que permita descubrir relaciones entre la información. Un excelente ejemplo de cómo funcionara la navegación es el sitio de Wikipedia.

• La Enciclopedia Galáctica presenta un alto número de lecturas contra un bajo número de escrituras (prácticamente 0).

• Para el año 23651, se han escrito:

o 4 billones de libros con una media de 200 páginas.

o 1 billón de artículos científicos con una media de 10 páginas.

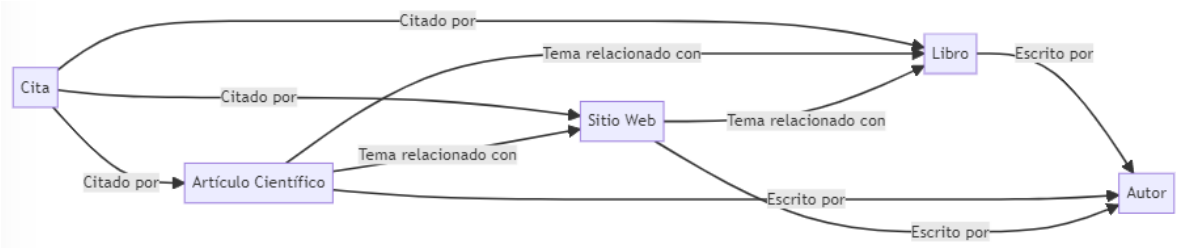
o 20 billones de sitios web con una media de 10 páginas cada uno.

En su calidad de líder técnico, usted debe presentar una propuesta para dar respuesta a las siguientes preguntas:

1. ¿Qué motor de base de datos utilizaría para implementar la navegación entre distintos elementos de información? ¿Es necesario que este motor de base de datos contenga todo el elemento de información o solo palabras clave que permitan establecer relaciones? Justifique su respuesta mediante la elaboración de un pequeño modelo de datos y las relaciones que establecería entre los diferentes elementos de información, lo más importante es garantizar una navegación y que permita descubrir relaciones. (20 pts)

Para proporcionar la navegación entre distintos elementos de información, implementaría un motor de base de datos orientado a grafos que funcione en combinación de un motor de búsqueda de texto completo. Neo4j es reconocido por liderar en bases de datos de grafos, con este se pueden indexar nodos y relaciones con propiedades en string. Me parece una buena opción debido a su sistema de consultas avanzadas que permite descubrir más relaciones y patrones y su integración con ElasticSearch para búsqueda de texto completo. No es necesario que contenga todo el contenido completo, se pueden usar palabras clave y metadatos que permitan establecer las relaciones y realizar búsquedas de forma eficiente.

A continuación, se presenta un modelo de datos simple. Los nodos serían los libros, artículos científicos, sitios web, autor y citas, estos incluyen toda la información referente al título, fecha de publicación, descripción, tema, citas e identificador.



2. ¿Qué motor de base de datos utilizaría para almacenar los elementos de información y garantizar full text search? Justifique su respuesta comentando: (20 pts)

Para almacenar los elementos de información y garantizar el full text search usaría ElasticSearch.

a. Capacidad del motor para implementar full text search.

ElasticSearch es reconocido por su capacidad de búsqueda de texto completo sumamente eficiente y útil. Este esta diseñado para realizar búsquedas rápidas y precisas en volúmenes sumamente altos de datos como lo es la Enciclopedia Galáctica. El motor es distribuido y escalable, puede indexar y buscar texto en tiempo real.

b. Particionamiento o sharding de datos.

Elasticsearch permite particionar datos y distribuirlos en varios nodos o servidores. Esto le da acceso a escalabilidad horizontal al sistema para agregar nuevos nodos y distribuir de forma eficiente la carga de trabajo. Así ElasticSearch puede aprovechar la distribución de carga para mejorar el rendimiento y las respuestas.

c. Representación de elementos de información en la base de datos (tablas, documentos, collections, etc.)

ElasticSearch se destaca por su modelo de datos flexible basado en documentos JSON. No se requiere una estructura ya definida para los documentos, esto facilita la adaptación a los diferentes tipos de información que tiene la Enciclopedia Galáctica, cada uno con sus atributos específicos y la posibilidad de agregar más.

3. Describa la forma en la cual combinaría los dos motores anteriores (navegación y full text search) para crear un sistema simple de búsqueda y navegación de información similar al que tiene el sitio Wikipedia donde se busca un elemento de información y nos podemos mover entre términos. (5 pts)

Primero sería necesario usar ElasticSearch para indexar los datos de toda la Enciclopedia Galáctica. Luego usaría Neo4j para almacenar la estructura de grafo y las relaciones entre los elementos de información usando palabras clave y metadatos. Consiguiente establecería una conexión entre ambos motores para permitir una búsqueda y navegación integrada. En el momento que se realice una consulta de búsqueda full text en ElasticSearch y se obtengan los resultados relevantes, se utilizarán los identificadores únicos creados para que se realicen consultas en Neo4j y obtener información adicional sobre relaciones entre elementos. Para la navegación entre elementos se usa Neo4j, se puede presentar en el frontend enlaces en azul a base de las palabras clave y metadatos en la base de datos que permitan al usuario explorar y moverse entre diferentes páginas de información de forma sencilla siguiendo las relaciones establecidas en la Enciclopedia Galáctica. También existe una integración entre ambos los resultados de búsqueda de full text y los datos de navegación de ElasticSearch para ver tanto resultados relevantes como enlaces adicionales a elementos relacionados.

4. ¿De qué forma garantizaría alta disponibilidad de las bases de datos? (5 pts)

Para garantizar la disponibilidad de la base de datos, configuraría la replicación en ambas bases de datos. Haría múltiples copias de los datos distribuidos en diferentes servidores a través de la galaxia. En caso de que algún servidor fallé, los demás pueden asumir la carga y mantener la disponibilidad de los datos. Cada centro de datos donde se encuentren los servidores serán llamados “Fundación”.

5. ¿Cómo podría garantizar que las búsquedas siempre tengan un tiempo de respuesta constante? (5 pts)

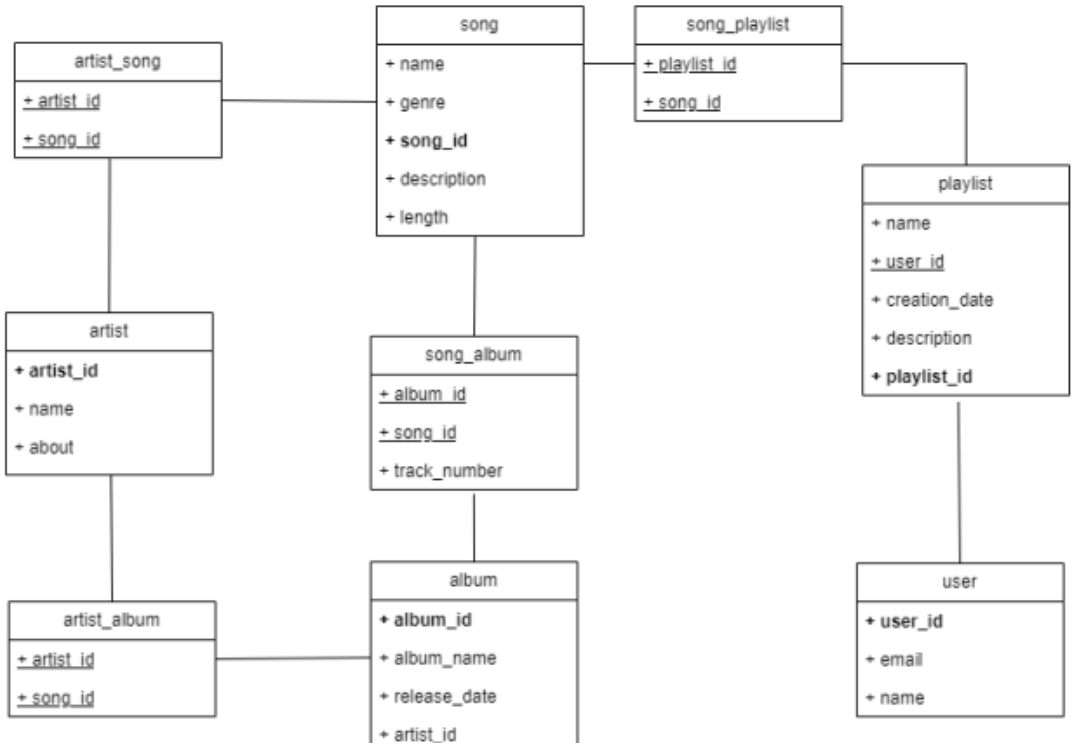
Ya que se están usando diferentes servidores o nodos en toda la galaxia en centros de datos, usaría escalabilidad horizontal para distribuir la carga de trabajo. Esto significa agregar más nodos o réplicas para distribuir las consultas y evitar la sobrecarga. Esto permitiría mejorar el rendimiento y el manejo del alto volumen de consultas para que el tiempo de respuesta se rápido y constante. También me centraría en hacer una indexación adecuada con los campos relevantes y optimizar las consultas lo más posible.

6. ¿Cómo el uso de caches y localidad podría mejorar el rendimiento del sistema? (5 pts)

El uso de caches y localidad puede mejorar el rendimiento del sistema al almacenar en memoria los datos más frecuentemente utilizados alrededor de la galaxia. Esto evita constantes accesos costosos a la base de datos, reduciendo la latencia y disminuyendo el tiempo de respuesta de las consultas. Además, complementan en escalabilidad al distribuir la carga y reducir presión sobre recursos principales. Conforme a localidad, sí se organiza el cache de forma correcta se pueden acceder datos relacionados de forma contigua al cache y así se reduce la necesidad de acceder a múltiples ubicaciones en memoria.

Pregunta 2 (10 pts)

El siguiente diagrama representa una versión simplificada de un sistema de reproducción de música que utiliza una base de datos relacional:



Este sistema tiene varios vicios o problemas de normalización, así como el grave problema de que no tiene definidos índices, en conjunto esto ha causado que se esté experimentando muchos timeouts y la solución convencional de agregar más hardware se ha vuelto insostenible. Luego de un estudio del workload de la base de datos, se llegó a las siguientes conclusiones:

• Es necesario definir algunos índices fuera de los que son definidos automáticamente mediante llaves primarias y foráneas.

• Un motor de base de datos relacional no parece ser el más adecuado para el problema.

• El patrón de uso es muchas lecturas contra pocas escrituras.

Mediante los logs de acceso y los logs de slow queries, se ha encontrado que los siguientes queries son los más usuales y problemáticos en tiempo que tardan en ejecutarse:

SELECT name FROM artist WHERE name like ‘%{text}%’

SELECT name FROM album WHERE name like ‘%{text}%’

SELECT a.name as artist_name, al.name as album_name, s.name, s.genre, sa.track_number,

s.length, s.description FROM artist a INNER JOIN artist_song as ON a.artist_id = as.artist_id INNER JOIN song s ON as.song_id = s.song_id INNER JOIN song_album sa ON s_song_id = sa.song_id INNER JOIN album al ON sa.album_id = al_album_id WHERE a.name = ‘%{name}%’ AND al.name = ‘%{name}%’ and s.name =‘%{name}%’

Como administrador o administradora de bases de datos, elabore respuestas a las siguientes preguntas:

1. ¿Qué índices definiría para aumentar la velocidad de todo el sistema? Tome en cuenta todos los tipos de índices estudiados en el curso. (3 pts)

De primero haría un índice de texto completo de la columna name a la tabla artista, esto para acelerar las consultas donde haya un SELECT en la tabla artista que busca por nombre. También haría uno similar pero este índice sería de la columna name a la tabla álbum. Este sería de texto completo y mejoraría consultas donde haya un SELECT en la tabla album que busca por nombre. Por último, añadiría un índice compuesto en las columnas name, artista, álbum y song. Este es debido a que en el tercer query se hace una búsqueda por nombre en estas tres tablas. Esto ayudaría el rendimiento de todos los JOIN que involucra este query en esas tablas.

2. ¿Qué base de datos SQL o NoSQL recomendaría para reemplazar la base de datos actual? Justifique su respuesta. (3 pts).

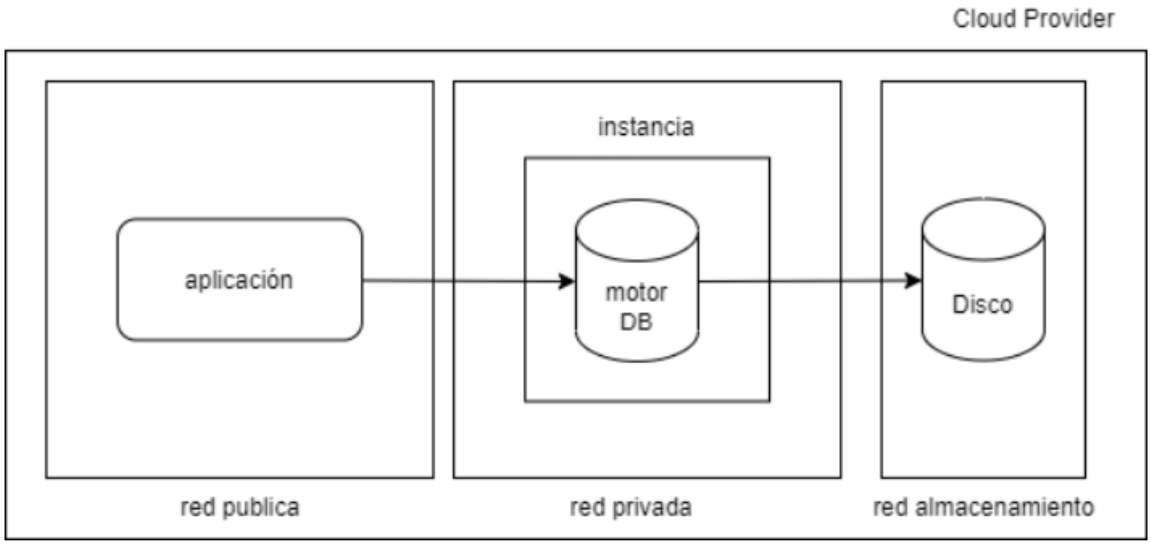
Tomando en cuenta motor de base de datos relacional no parece ser el más adecuado y que el patrón de uso es muchas lecturas contra pocas escrituras, mi recomendación sería usar una base de datos NoSQL. En este caso MongoDB me parece la mejor opción debido a su flexibilidad tanto en búsquedas avanzadas como en esquema y su alto rendimiento en entornos de lectura intensiva.

3. ¿Existirá alguna otra forma de mejorar el rendimiento de la base de datos relacional en especial para la tercera consulta? Comente. (4 pts)

Sí existe otra forma de mejorar el rendimiento de la base de datos relacional en la tercera consulta. Se podría usar un sistema de caché de consultas para almacenar los resultados de esta consulta complicada que es más frecuente. Así se evita la necesidad de ejecutarla en la base de datos todas las veces.

Pregunta 3 (20 pts)

En los últimos 15 años, la forma en la cual se mantienen e instalan servidores de bases de datos ha cambiado considerablemente, la aparición del Cloud ha proporcionado muchas ventajas para la instalación y mantenimiento, pero ha inducido nuevos problemas de seguridad y nuevas soluciones, en el siguiente diagrama se muestra una arquitectura típica de una base de datos en un Cloud Provider:



Tomando como referencia el diagrama anterior, ¿Cuáles son las buenas prácticas en términos de seguridad que se deben seguir cuando se instala un motor de base de datos en el Cloud? Fundamente su respuesta hablando de la seguridad de cada uno de los componentes que se exponen en el diagrama.

Para la red pública sería importante implementar firewalls y reglas establecidas para restringir el acceso a la aplicación únicamente para los IP autorizados y filtrar el tráfico de información. Es importante encriptar la comunicación entre la aplicación y la base de datos usando protocolos seguros (https). También implementar mecanismos de autenticación y autorización estables para controlar el acceso a la aplicación.

Para la red privada es importante habilitar un registro y mantener un monitoreo constante de las actividades. Sí pasa algo fuera de la normal se puede detectar como actividades maliciosas o posibles intrusiones no autorizadas al sistema. También es necesario crear firewalls en la red privada que permitan mantener las políticas de seguridad del motor. Es importante conocer bien el sistema y tener al tanto las vulnerabilidades conocidas. Configurar grupos de seguridad o listas de control de acceso (ACL) para restringir el acceso a la instancia del motor de base de datos desde la red pública es una práctica esencial.

Para la red almacenamiento me parece esencial encriptar los datos sensibles almacenados en el disco para proteger la confidencialidad en caso de un acceso físico o de infraestructura no autorizado. También es importante realizar copias de seguridad para garantizar disponibilidad y protegerse de los casos de pérdida y corrupción de datos.

Pregunta 4 (10 pts)

La Observabilidad es una gran herramienta que nos permite tener una visión en el tiempo de la forma en la cual se comportan sistemas computacionales, estos sistemas hacen uso extensivo de bases de datos de series de tiempo, una de las más utilizadas es Prometheus, pero existen soluciones que utilizan otras bases de datos o motores de búsqueda como Elasticsearch u OpenSearch. Como ingeniera o ingeniero a cargo de los sistemas de Observabilidad de una

empresa, se le ha solicitado dar respuesta a las siguientes preguntas, con el fin de determinar la estrategia que seguirá la empresa en términos de Observabilidad en los siguientes años.

1. ¿Por qué las bases de datos de series de tiempo son tan utilizadas en soluciones de Observabilidad? Realice un análisis desde el punto de vista de la naturaleza de los datos que se recolectan. (2 pts)

Los datos recolectados en observabilidad se basan en el seguimiento y registro de eventos continuos y métricas de tiempos. Estos al ser organizados secuencialmente y con marca de tiempo los convierten en datos de series de tiempo. Las bases de datos de series de tiempo están especialmente diseñadas para almacenar y analizar este tipo de datos secuenciales.

2. ¿Es posible utilizar BigTable como una base de datos de series de tiempo que se pueda utilizar como parte de una solución de Observabilidad? Justifique su respuesta desde el punto de vista de la naturaleza de la base de datos. (2 pts)

Sí es posible usar BigTable como una base de datos de series de tiempo para utilizar como parte de una solución de observabilidad. Google ofrece opciones de diseños de esquemas para la colección de datos de series de tiempo en BigTable. En los dos patrones presentados las filas pueden ser buckets de tiempo o marcas de tiempo, dependiendo de como se quieran almacenar los datos. Igual es importante destacar que a pesar de que se puede, BigTable no está diseñada específicamente para manejar datos secuenciales y no ofrecerán las funcionalidades especializadas para consultas y análisis que hay en una base de datos de series de tiempo, lo cual la hace poco adecuada para una solución de observabilidad.

3. Suponiendo que tenemos una solución de Observabilidad que utiliza Elasticsearch, ¿Cómo podemos ahorrar dinero con información histórica? (2 pts)

Para ahorrar dinero con información histórica en una solución de Observabilidad que utiliza Elasticsearch, es importante establecer políticas de retención de datos claras, definiendo el período de tiempo necesario para el análisis. También comprimir los datos almacenados para reducir el tamaño de almacenamiento y realizar snapshots periódicos y guardarlos en un almacenamiento de respaldo para liberar espacio en el clúster. Estas medidas permiten optimizar los recursos y reducir costos sin comprometer la accesibilidad y capacidad de consulta de los datos históricos.

4. Comente las ventajas y las desventajas de utilizar un servicio de Observabilidad on-premise (por ejemplo, Prometheus y Grafana) vs un Managed Service (como Datadog), justifique su respuesta con la experiencia obtenida en la tarea corta 1 de este curso. (4 pts)

La ventaja principal que se notó en usar un servicio de Observabilidad on-premise es que dentro de la organización hay un control total sobre todos los aspectos de configuración, seguridad y personalización de la plataforma. En un Managed Service la organización depende completamente del proveedor y los servicios que ofrezca. Otra de las ventajas de on-premise es que a diferencia de los Managed Service, no hay costos recurrentes por el servicio, por lo que se generan ahorros significativos. Entre las desventajas de on-premise se da que la complejidad de implementar la solución de monitoreo es mayor, requiere más experiencia, estudio y dedicación por parte del equipo para configurar y mantener. Los Managed Service son de una implementación más rápida y

fácil, con sistemas de mantenimiento simplificados. También en el on-premise se hace un mayor esfuerzo debido a que posee muchas más menos opciones de automatización integrada y soporte que los Managed Service sí suelen tener. Estos aspectos permiten que en los Managed Service se pueda centrarse más desde el principio en observabilidad y menos en implementación, lo cual me parece la mayor desventaja de usar on-premise.

Referencias.

https://www.elastic.co/guide/en/elasticsearch/reference/current/full-text-queries.html

https://neo4j.com/docs/getting-started/data-modeling/

https://cloud.google.com/bigtable/docs/schema-design-time-series?hl=es-419

https://www.educba.com/datadog-vs-prometheus/

https://static.googleusercontent.com/media/research.google.com/es//archive/bigtable-osdi06.pdf

https://www.elastic.co/es/what-is/observability

https://www.elastic.co/es/blog/elastic-observability-clusters-upgrade-latest-release-save-money