Examen

Tecnológico de Costa Rica
Escuela de Ingeniería en Computación
Bases de Datos II (IC 4302)
Primer Semestre 2023

Datos del estudiante:

Gerald Núñez Chavarría - 2021023226

Instrucciones:

- Conteste todas las preguntas con el nivel mínimo y suficiente de detalle para demostrar su conocimiento del tema.
- No se evaluarán respuestas parciales o imprecisas.
- Es responsabilidad del estudiante garantizar que sus respuestas se entiendan, puede usar recursos como imágenes, diagramas, videos, etc.
- Si las respuestas no se entienden el profesor está en derecho de calificar con un 0 la respuesta.
- La nota máxima del examen es 100.
- El tiempo estimado para completar el examen en una clase presencial es de 120 minutos.
- El examen deberá ser entregado antes de las **05:00 pm del día 10 de junio del 2023**. El estudiante cuenta con más de 12 horas para elaborar el examen. La fecha en la cual se entregaría el examen fue notificada con más de 15 días de antelación y acordada con todo el grupo.
- Si el examen es entregado después de esta hora, no será revisado y se obtendrá una nota de 0.
- El examen deberá ser entregado al correo electrónico del profesor, debe seguir el formato especificado en el programa del curso.
- El nombre del archivo debe ser **ex.pdf**
- Cualquier indicio de copia será calificado con una nota de 0 y será procesado de acuerdo con el reglamento, esto incluye cualquier herramienta que genere textos mediante inteligencia artificial y cualquier producción parcial o total de algún documento sin su debido reconocimiento al autor o autora.
- Se puede utilizar cualquier recurso en Internet para elaborar sus respuestas, deben especificar referencias bibliográficas, se debe validar que sea una fuente confiable, herramientas de inteligencia artificial no se consideran una fuente confiable.
- Si la referencia bibliográfica NO es confiable el profesor está en derecho de calificar la respuesta con una nota de 0. Para verificar si la referencia el confiable, puede hacer las siguientes preguntas:
    - ¿Quién es el autor o autora?
    - ¿Cuál es el propósito de ese documento?
    - ¿Es posible que esté parcializado?
    - ¿Ha sido revisado o aprobado por expertos en ese campo de estudio?
- Las preguntas fuera del horario de clase se pueden hacer por medio de correo electrónico o al grupo oficial de Telegram, pueden darse retrasos en las respuestas a las preguntas, en especial las que se realizan a altas horas de la noche o madrugada, se recomienda realizar todas las consultas necesarias durante la clase del 9 de junio del 2023.
- El examen consta de 4 preguntas de desarrollo.

- **Es importante recalcar que las preguntas son de desarrollo, cada respuesta debe estar cuidadosamente desarrollada con explicaciones adecuadas.**
- El valor del examen es de un 10%.
- Es responsabilidad del estudiante completar todas las preguntas del examen, en caso de que se olvide responder alguna de ellas se obtendrá una nota de 0.

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:
    - 4 billones de libros con una media de 200 páginas.
    - 1 billón de artículos científicos con una media de 10 páginas.
    - 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

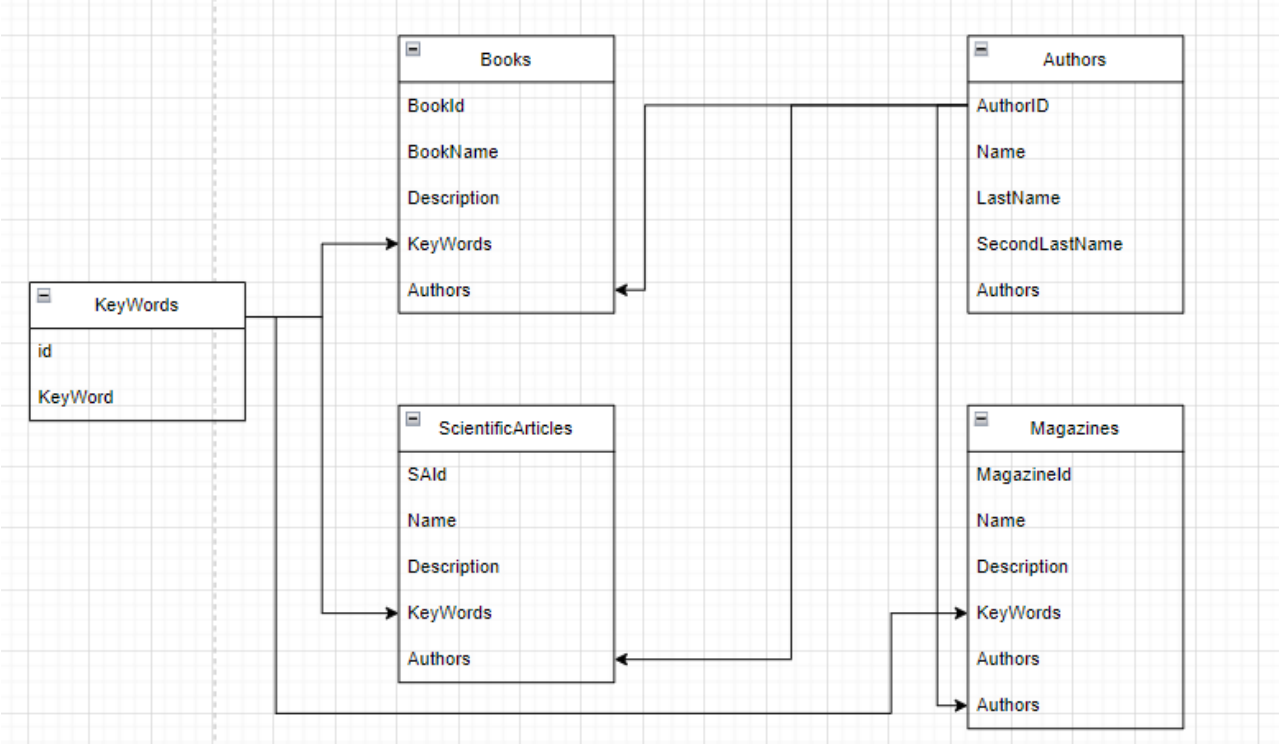
mediante la elaboración de un pequeño modelo de datos y las relaciones que establecería entre los diferentes elementos de información, lo más importante es garantizar una navegación y que permita descubrir relaciones. (20 pts)

Respuesta:

Utilizaría el motor de búsqueda de Elasticsearch ya que ofrece capacidades avanzadas de búsqueda y análisis de texto completo, lo que permite implementar la navegación que se solicitar para la Enciclopedia Galáctica.

Considero que no es necesario que el motor contenga todo el elemento de información, se pueden entre los elementos con palabras claves que permitan vincularse a otros elementos.

Diagrama de modelo de datos:



La idea es tener indexados palabras clave y autores por ejemplo en este caso, que ambas tienen relación con los libros, los artículos científicos y las revistas, para así mediante una palabra clave traer la información de ese X elemento. Se pueden implementar sinónimos y demás, pero el diagrama es solo para ejemplificar las relaciones que se pueden crear con el ejemplo de palabras clave y autores.

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)

   Respuesta:

   El motor de bases de datos seleccionado es MongoDB con su integración Atlas Search.

   a. Capacidad del motor para implementar full text search.
      Basándose en Apache Lucene, MongoDB integra Atlas Search que permite realizar full text search que incluyen funciones muy avanzadas e importantes como realizar ranking por el match, resaltar los términos dónde se hace match, realizar filtros y manejar facets, búsqueda de frases completas…

b. Particionamiento o sharding de datos.

La capacidad de sharding de MongoDB permite distribuir los datos de una colección entre múltiples fragmentos en un clúster fragmentado. Cada fragmento contiene un subconjunto de los datos y puede ser implementado como un conjunto de réplicas. El enrutador de consultas, conocido como mongos, actúa como intermediario entre las aplicaciones cliente y el clúster fragmentado.

Los servidores de configuración almacenan metadatos y ajustes de configuración. MongoDB utiliza una clave de fragmento para distribuir los documentos entre los fragmentos, y esta clave puede ser seleccionada al momento de fragmentar una colección. La capacidad de sharding brinda ventajas como la distribución de carga de trabajo, la capacidad de almacenamiento escalable y alta disponibilidad en caso de fallos en los conjuntos de réplicas. (Sharding — MongoDB Manual, n.d.)

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

En MongoDB se puede guardar los datos en documentos que son almacenados en una colección especifica, esto es ideal para la Enciclopedia Galáctica, ya que los diferentes elementos de información, como libros, artículos científicos y sitios web, pueden representarse fácilmente como documentos independientes en colecciones separadas. Cada documento puede contener los campos necesarios para almacenar las palabras clave, autores relacionados, contenido relacionado... Esto permite una representación eficiente de la información y un acceso rápido a través de consultas de búsqueda.

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)

Respuesta:

Utilizaría MongoDB para almacenar los datos. Luego utilizaría esos datos indexados a Elasticsearch para crear relaciones cercanas entre el contenido, como información del autor y contenido relacionado, generando links a las páginas cercanas. Después, cuando se realice una búsqueda, utilizaría la integración de Atlas Search para buscar devolver elementos relacionados a esa palabra o frase, resaltando los términos que hicieron match. Al seleccionar un elemento especifico se mostrará directamente la información completa trayéndola de MongoDB, pero además tendrá links a otros elementos relacionados que son indexados con Elasticsearch.

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

Respuesta:

Las réplicas son una buena opción ya que, si un nodo o servidos de cualquiera de los motores de bases de datos se cae, no se va a ver afectado el sistema, además se debe tener este sistema automatizado, es decir, la reasignación de réplicas o la elección de un nodo secundario como líder en caso de fallo del líder principal.

También utilizar clústeres para distribuir y balancear la carga de trabajo, asegurando que los servicios de búsqueda y almacenamiento de datos estén disponibles incluso en situaciones de alto tráfico o picos de demanda.

Implementa una capa de caché de lectura, para almacenar en memoria los resultados de consultas frecuentes y que cuando se necesite realizar consultar no se tenga que ir directamente hasta la base de datos, si no que estén disponibles rápidamente.

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

   Respuesta:

   Configurar índices adecuados y optimizar las consultas para aprovechar al máximo los índices, esto ayudará a manejar consultas muy rápidas independientemente de la cantidad de usuarios.

   Además, es importante utilizar técnicas de particionamiento de datos para distribuir la carga de trabajo de manera equitativa y mantener un tiempo de respuesta constante en todas las consultas.

   Se puede implementar el caché para tener resultados de consultas muy frecuentes en memoria, para devolver los resultados sin necesidad de ejecutarla.

   Diseñar escalabilidad horizontal para aumentar el número de nodos o servidores según sea necesario y así controlar un mayor volumen de consultas y mantener un tiempo de respuesta constante.

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

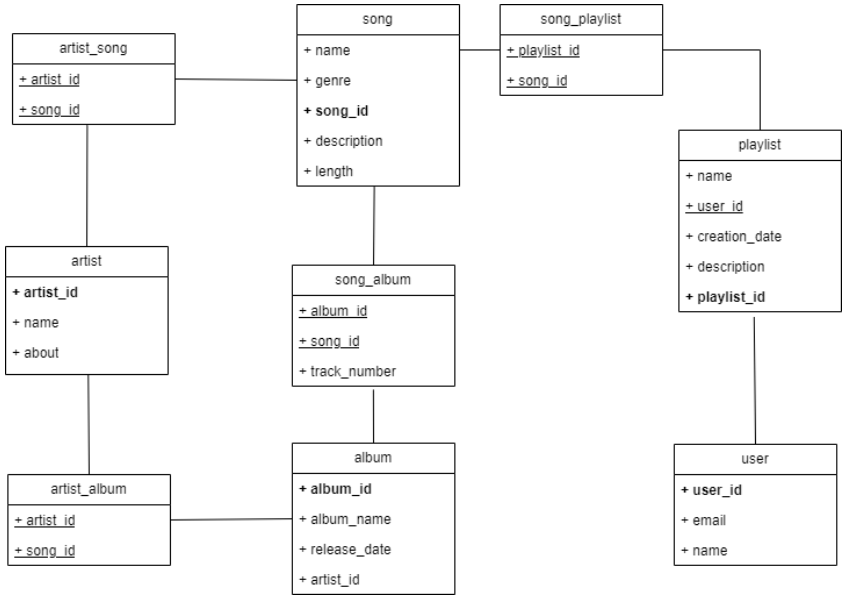
   Respuesta:

   La implementación de una caché de consultas permite almacenar en memoria los resultados de las consultas más frecuentes, evitando así la necesidad de ejecutar la consulta cada solicitud. Esto se traduce en un tiempo de búsqueda y recuperación de datos considerablemente reducido, ya que la información ya está disponible en la caché y puede ser entregada de manera rápida, además en una reducción de las consultas en la base de datos, evitando una sobre carga.

   La localidad de los datos es otro aspecto importante para mejorar el rendimiento del sistema. Si se organizan y relacionan los datos de acuerdo con su localidad, entonces se minimiza la necesidad de buscar y acceder a múltiples ubicaciones dispersas, lo que resulta en una mayor eficiencia y un menor uso de recursos.

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}%'

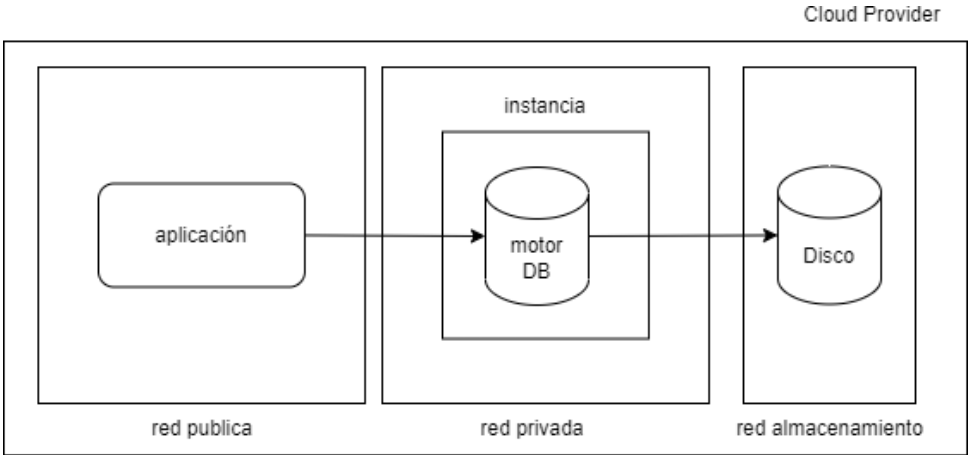
Como administrador o administradora de bases de datos, elabore respuestas a las siguientes preguntas:

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

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

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.

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.

- ¿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)

  Respuesta:

  Cuando se realiza observabilidad, se van a recolectar datos cómo el uso del disco, las conexiones a la base de datos, si las solicitudes son de escritura, lectura o ambas, queries por segundo… todos estos datos recolectan de manera lineal durante una serie de tiempo (dos horas por ejemplo), por lo tanto, se adaptan justamente a las bases de series de tiempo, ya que están hechas para almacenar estos tipos de datos.

- ¿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)

  Respuesta:

  Se puede dieñar una base de datos de BigTable con el objetivo de guardar datos de series temporales lo que le permite a una base de datos como BigTable, aunque no sea naturalmente para, servir como una herramienta de observabilidad. (Google Cloud, n.d.)

  Claramente no tiene las mismas funciones especializadas que bases de datos dedicadas, pero puede ser una opción viable.

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

- 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)

  Respuesta:

  Utilizar un servicio de observabilidad on-premise ofrece ventajas tener el control total sobre la configuración y personalización de la infraestructura de observabilidad, lo que permite adaptarlo a las necesidades. Además, permite tener mayor privacidad de datos y controlar los costos generados por el servicio.

  Manejar un servicio propio de observabilidad genera costos de implementación en tiempo y aprendizaje de habilidades técnicas del personal para poder configurarlo y además darle el mantenimiento necesarios, como actualizaciones, administración de hardware…

Por otro lado, un servicio administrado provee ventajas como la facilidad de uso y una configuración rápida, ya que los proveedores se aseguran de que la configuración sea muy intuitiva para sus usuarios. Una vez instalado, el proveedor se encargada del mantenimiento, lo cual libera esa preocupación del usuario.

Entre las ventajas se puede mencionar la dependencia y confianza en un proveedor, ya que, si este presenta problemas, el usuario no puede hacer nada. Tiene un mayor costo que un servicio on-premise.

Referencias

Google Cloud. (n.d.). Diseño de esquemas para datos de series temporales. *Google Cloud*. https://cloud.google.com/bigtable/docs/schema-design-time-series?hl=es-419

Sharding — MongoDB Manual. (n.d.). https://www.mongodb.com/docs/manual/sharding/