Base de Datos II (IC4302) – Semestre 2, 2023

Examen

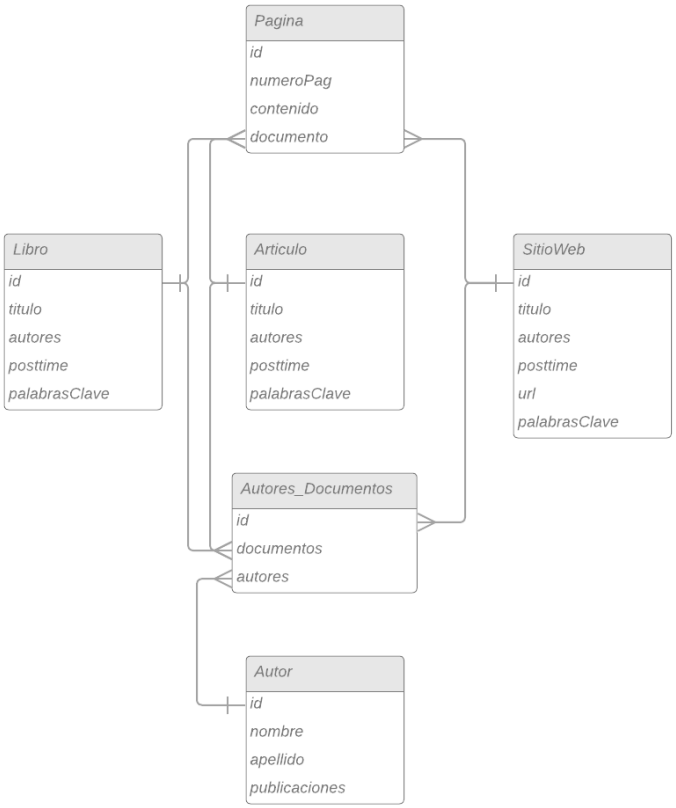
David Suárez Acosta – 2020038304

Pregunta 1 (40 pts)

1. **¿Qué tipo de base de datos (SQL o NoSQL) 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, diagrama y las relaciones que establecería entre los diferentes elementos de información, lo más importante es garantizar una navegación y que permita descubrir relaciones. (10 pts)**

El tipo de base de datos que usaría sería una base de datos SQL ya que, como por el momento el enfoque es la navegación entre distintos elementos de información, resulta mejor la estructura y relaciones de una base de datos relacional. Con solo que el motor de base de datos contenga palabras claves se pueden establecer relaciones ya que se va a permitir “full text search” más adelante, por lo que no es necesario que se almacene todo el elemento de información.

A continuación, se muestra un modelo de datos SQL que se puede usar para la Enciclopedia Galáctica:



2. **¿Qué motor de base de datos utilizaría para almacenar los elementos de información y garantizar full text search? Justifique su respuesta comentando: (10 pts)**
    a. **Capacidad del motor para implementar full text search.**
    b. **Particionamiento o sharding de datos.**
    c. **Representación de elementos de información en la base de datos (tablas, documentos, collections, etc.)**
    d. **Localidad de datos.**

Para este caso usaría MongoDB, el cual es un motor de base de datos NoSQL, esta selección se debe a que:

- MongoDB implementa de forma completa el "full text search". (https://www.mongodb.com/basics/full-text-search)
- MongoDB permite el particionamiento de datos con "Chunks". (https://www.mongodb.com/docs/manual/core/sharding-data-partitioning/)
- MongoDB usa documentos JSON Binary para el almacenamiento de datos. (https://www.mongodb.com/docs/manual/core/document/)

- MongoDB está distribuida en diferentes ubicaciones a nivel global. (https://www.mongodb.com/products/platform/cloud)

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. (4 pts)**

Para combinar la navegación con SQL con el "full text search" de MongoDB se pueden llenar ambos motores con datos, el SQL con el modelo establecido en el punto 1. y MongoDB con información que le permita hacer el "full text search", luego de esto se deben sincronizar los datos de ambos motores para crear una conexión entre estos (por esta razón se dejó el campo de "Palabras Claves" en el modelo de la base SQL) y por último crear un sistema que realice búsquedas por "full text search" y que, en caso de querer ver la información completa, se utilicen consultas al motor SQL.

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

Se puede garantizar alta disponibilidad de las bases de datos si se distribuyen los datos con replicación por diferentes ubicaciones de la galaxia, así no importa en donde se esté siempre va a haber un servidor cercano que permita la alta disponibilidad.

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

La replicación de datos mencionada en el punto 4. también permite que las consultas, que son casi en su totalidad de lectura, se puedan realizar al servidor más cercano, de esta forma no se sobrecarga un servidor específico. Para mejorar el tiempo de respuesta también se pueden incluir índices que permitan el acceso rápido a los datos.

6. **¿Cómo el uso de caches y localidad podría mejorar el rendimiento del sistema? (4 pts)**

Se pueden utilizar sistemas de cache, los cuales pueden estar distribuidos por la galaxia, en los cuales se almacenen las respuestas de consultas frecuentes que se realizan a las bases de datos. Con respecto a la localidad, se podrían ubicar servidores de almacenamiento de datos similares (como por ejemplo todos los artículos relacionados al tema X) en ubicaciones cercanas para evitar que se tengan que hacer consultas que duren mucho tiempo para temas similares. Todo esto podría mejorar el rendimiento del sistema.

7. **¿Describa como realizaría el procesamiento de los datos y su carga en las bases de datos? Tome en cuenta el volumen, la variedad y velocidad de carga de los datos (Big Data) (8 pts)**

Para realizar el procesamiento y carga de los datos se pueden utilizar procesos ETL, los cuales son procesos diseñados para extraer, transformar y cargar datos de tipo Big Data en sistemas, como en este caso el sistema de base de datos establecido en los puntos anteriores que use una base SQL para la navegación y MongoDB para el “full text search” y que reciba cantidades tan grandes de datos (https://www.iebschool.com/blog/que-son-los-procesos-etl-big-data/).

Al mismo tiempo, ya que estas operaciones de escritura serían tan ocasionales, se puede utilizar la programación asincrónica para la carga de datos, esto debido a que de esta forma el proceso no afecta las operaciones principales del sistema, las cuales serían posiblemente operaciones de lectura.

Pregunta 2 (40 pts)

Dar una solución detallada de cómo podría mejorar el rendimiento de la base de datos actual, reduciendo el downtime al mínimo, esto permitirá ganar tiempo para dar una solución mucho más duradera con la mínima afectación a los usuarios. (5 pts)

Para mejorar el rendimiento y reducir el downtime se puede:

- Realizar un escalamiento horizontal y crear más servidores que dividan la base de datos, ya que actualmente todo se maneja en la casa de uno de los fundadores.
- Implementar un sistema de cache para manejar las consultas repetitivas.
- Usar índices para manejar las consultas similares.

Dar una recomendación detallada de que tipo de base de datos se debería utilizar para abordar este problema, además debe recomendar algunas de las bases de datos SQL o NoSQL estudiadas durante el curso tanto en lecturas, así como las utilizadas en proyectos o ejemplos en clase. Tome en cuenta que sería posible utilizar más de una base de datos para optimizar el almacenamiento de los datos de la tabla post, amigos y usuario, tome en cuenta que tan fácil es escalar la base de datos en su recomendación, debe dar prioridad a servicios managed services, SaaS y PaaS, no olvide la localidad y naturaleza de los datos. (20 pts)

Recomiendo el uso de una base de datos relacional, específicamente se puede usar la base de datos PostgreSQL para optimizar el almacenamiento de los datos de las diferentes tablas. El uso de PostgreSQL permite la escalabilidad horizontal, como la mencionada en el punto anterior para mejorar el rendimiento de la base.

PostgreSQL se puede adaptar a servicios “managed services” con el uso de Amazon RDS para PostgreSQL como una base de datos “managed” (https://aws.amazon.com/es/rds/postgresql/), al mismo tiempo tanto SaaS como PaaS pueden utilizarse con PostgreSQL ya que puede ser utilizado para aplicaciones basadas en la nube (SaaS) y puede ser usado como el motor de base de datos para aplicaciones de plataforma que se gestionan en la nube (PaaS).

Por último, la localidad de los datos puede variar dependiendo de donde se utilice más la aplicación, si la aplicación llega a expandirse a nivel global habría que hacer un escalado horizontal a nivel internacional para que la localidad de los datos este esparcida por el mundo. Debido a que los datos son tablas, lo mejor es el uso de una base de datos SQL como lo es PostgreSQL.

Comente acerca de que tan conveniente es mantener la base de datos actual en la casa de uno de los fundadores, comparado con mover ésta algún Cloud Provider como AWS. (5 pts)

Mantener la base de datos actual en la casa de uno de los fundadores no es muy conveniente en comparación con moverlo a un Cloud Provider ya que:

- El espacio es limitado y esto puede llegar a ser un problema como se esta viendo, ya que el rendimiento baja y se limita el crecimiento de la base. Si se usara un Cloud Provider no existirían estos problemas ya que los Cloud Provider son escalables.
- Puede haber problemas de disponibilidad ya que es muy fácil que ocurran problemas en un lugar individual como lo es la casa del fundador, estos problemas pueden ser por ejemplo que se vaya la luz o se vaya el internet.
- Es muy poco seguro que se mantengan en una casa, en comparación de que se guarden en bases de datos de verdad protegidas por diferentes barreras físicas.

La única ventaja que podría tener mantenerla en la casa del fundador es que se evita pagar el Cloud Provider pero al final de cuentas este costo puede llegar a ser necesario si se quiere un buen sistema con buen rendimiento y otras ventajas que tienen los Cloud Provider.

Basándose en el funcionamiento de un índice invertido el cual fue estudiado en clase y es utilizado por motores como Elasticsearch y el concepto de Natural Languaje Processing (NLP) llamado Stemming, comente ¿Cómo se podría reducir el memory footprint de la base de datos actual? (10 pts)

Para reducir el “memory footprint” se puede:

- Eliminar palabras muy comunes y de poco valor como lo son las palabras “y”, “o”, “el”, etc.
- Se pueden usar técnicas Stemming para que todas las palabras se conviertan en su raíz y así eliminar más palabras, por ejemplo, se pueden convertir “corrió” y “corriendo” a “correr”.
- Se pueden usar índices invertidos para almacenar las palabras importantes, eliminando las mencionadas anteriormente, de esta forma se reduce el memory footprint actual.

Pregunta 3 (10 pts)

¿Cómo afectan los índices en el rendimiento de las bases de datos relacionales? Enfoque su respuesta tanto en como benefician el rendimiento así la forma en la cual lo impactan de forma negativa. (5 pts)

Los índices pueden beneficiar el rendimiento de las bases de datos relacionales ya que mejoran la velocidad de las búsquedas de datos, esto se debe a que los índices funcionan como un filtro para los datos.

Los índices también pueden afectar negativamente ya que usan espacio de disco, es decir que si se crean muchos índices se consume mucho espacio. Los índices también pueden afectar el rendimiento si se crean índices que no sirven o son innecesarios, de esta forma solo se consume almacenamiento por gusto.

Suponiendo que el hardware no es un problema (se puede comprar cuanto se necesite), ¿Podemos crear cuantos índices queramos o estos no tendrán mayor impacto en el rendimiento? (5pts)

Si se crean muchos índices puede llegar a impactar al sistema de forma negativa, ya que cuando se crean muchos índices se consumen los recursos del sistema y esto puede llevar a la competencia de recursos con el Sistema Operativo, lo cual puede llegar a hacerse un problema más grande.

Pregunta 4 (10 pts)

¿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)

Las bases de datos de serie de tiempo consisten en un sistema para almacenar datos tomados en intervalos de tiempo, funcionan como una especie de historial y es por esta razón que son tan buenos para soluciones de Observabilidad, ya que con estas bases de datos se pueden observar los diferentes eventos que han ocurrido en un tiempo y se pueden buscar patrones o tendencias de los 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)

Sí se podría utilizar BigTable como una base de datos de serie de tiempo ya que BigTable esta hecha para manejar cantidades muy grandes de datos como las que se producen con las bases de datos de serie de tiempo, por lo que se podría adaptar a este tipo de ingreso de datos y funciona como una solución de Observabilidad. (https://cloud.google.com/bigtable?hl=es)

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

Se podría crear una configuración que defina cuanto tiempo van a retenerse ciertos datos, de esta forma se evita que se gasten muchos recursos almacenando datos que ya fueron analizados y ya no es conveniente que sigan almacenados, por ejemplo si se analizan los datos de Instagram a lo largo del año, podría configurarse el Elasticsearch para que solo mantenga los datos de los últimos 3 meses.

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)

Al usar servicios de Observabilidad on-premise se puede manejar el servicio como se quiere, es decir que se pueden crear por ejemplo rúbricas específicas o en general configurar el servicio como se necesite; es muy flexible a diferencia de un Managed Service.

Otra ventaja es que se pueden ahorrar los costos de un Managed Service el cual requiere de gestión externa.

Las desventajas de usar servicios on-premise es que hay que saber como funciona la configuración de las herramientas y como se deben usar para lograr obtener cierto resultado, esto es mucho más fácil en un Managed Service ya que uno no se encarga de la configuración.