Examen

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

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 11 de noviembre 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 (nombre **ex1**).
- El nombre del archivo debe ser **ex1.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.
- En caso de que exista duda sobre alguna respuesta, el profesor podrá solicitar una defensa de esta, donde se comprobara que la respuesta fue elaborada por la persona estudiante y que se entienden los conceptos expuestos, si la defensa no es satisfactoria, se obtendrá una nota de 0.
- 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 10 de noviembre 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 (40 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:

- Las tecnologías 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, eso no quiere decir que las tecnologías utilizadas por Wikipedia son aptas para dar respuesta a esta pregunta.
- 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 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)
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.
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)
4. ¿De qué forma garantizaría alta disponibilidad de las bases de datos? (4 pts)
5. ¿Cómo podría garantizar que las búsquedas siempre tengan un tiempo de respuesta constante? (5 pts)
6. ¿Cómo el uso de caches y localidad podría mejorar el rendimiento del sistema? (4 pts)
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)

Pregunta 2 (40 pts)

Time No More es la red social más pequeña jamás creada, es conocida como ***one word social network (OWSN)***, este término fue acuñado por un profesor mientras diseñaba un examen, el funcionamiento es muy simple, como en cualquier otra red social las personas pueden tener amigos y postear mensajes, pero en este caso son simples mensaje con una palabra.

En sus primeros 2 meses, Time No More logró reunir cerca de 500 mil usuarios, con un volumen de queries de escrituras (nuevos posts) de 5 posts por segundo y un volumen de queries de lectura de 100 queries por segundo.

Originalmente esta OSWSN fue creada en una base de datos relacional, se puede asumir que es un motor MariaDB standalone en se encuentra la casa de uno de los fundadores, esto está causando muchos problemas de rendimiento y se estima que en menos de 4 meses si el crecimiento se mantiene como en los dos primeros meses la base de datos colapsará.

Esta red social, aunque simple se ha hecho popular, ya que permite entre otras cosas, mostrar el estado de ánimo, exponer deseos, aspiraciones, y otros, todo en una sola palabra, lo cual es sumamente conveniente para hacer análisis de mercadeo y además simplifica los modelos de inteligencia artificial.

También entre la juventud se hizo muy popular enviar mensajes palabra por palabra o hasta construir historias, donde cada persona escribe una palabra con un hashtag (máximo 1) y mediante el tiempo de inserción en la base de datos, se va construyendo una historia, esto ha sido utilizado por grupos con malas intenciones para enviar mensajes inapropiados o coordinar grupos de odio, que normalmente evitan enviar el hashtag para evitar ser descubiertos, al ser una base de datos relacional y al ser una característica muy utilizada, la base de datos relacional no está funcionando correctamente ya que los queries duran mucho tiempo.
Cada registro de la base de datos Time No More, tiene el siguiente formato:

Tabla **post**:

id	timestamp	hashtag	palabra	id_usuario	latitud	longitud

Tabla **usuario**:

id	nombre	apellidos	teléfonos	país	estado	zip

Tabla amigos:

id_usuario1	id_usuario2

En calidad de CTO de la compañía Timeouts No More, ustedes han sido contactados o contactadas para ayudar a Time No More e intentar solucionar el problema, por esta razón debe realizar las siguientes tareas:

- 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)
- 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)
- 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)
- 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)

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)
- 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)

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)
- ¿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)
- 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)