Fundamentos de bases de datos | **Facultad de Ciencias UNAM**
Tarea 01 | **Jurassic Team**
08/03/2022

Preguntas

1. **Conceptos generales**

   a. ¿Qué ventajas y desventajas encuentras al trabajar con una base de datos?

   b. ¿Qué es la independencia de datos? ¿Cuál tipo de independencia de datos es más difícil de lograr? Justifica tu respuesta.

   c. Explica la diferencia entre los esquemas externo, interno y conceptual. ¿Cómo se relacionan estas diferentes capas de esquemas con los conceptos de independencia de datos lógica y física?

   d. Investiga qué papel juegan los analistas de bases de datos, diseñadores y desarrolladores de bases de datos en la construcción de un sistema de bases de datos.

   e. Describe las relaciones que existen entre una base de datos y un Sistema Manejador de Bases de Datos.

   f. Entrevista a algún usuario de sistemas de bases de datos, ¿Qué características de SMBD encuentran más útiles y por qué? ¿Qué instalación(es) de SMBD encuentran más menos complicada y por qué? ¿Cuáles perciben estos usuarios que son las ventajas y desventajas de un SMBD?

   g. Supón que deseas crear un sitio de videos similar a YouTube. Considera cada uno de los puntos enumerados en el documento "Purpose of Database Systems", como desventajas de administrar los datos en un sistema de procesamiento de archivos. Discute la relevancia de cada uno de los puntos indicados con respecto al almacenamiento de datos de los videos: título, el usuario que lo subió, la fecha de carga, las etiquetas, qué usuarios lo vieron, cantidad de "Me gusta", entre otros.

   h. Indica las principales responsabilidades de un Sistema Manejador de Bases de Datos. Para cada responsabilidad, indica qué problemas que surgirían si la responsabilidad no se cumpliera. Justifica en cada caso tu respuesta.

   i. Asumiendo que una base de datos es un lugar donde se almacenan datos de forma sistemática y que la información se obtiene al consultar los datos entonces, un diccionario puede considerarse como una base de datos. Imagina que vas a buscar el significado de la palabra Luminiscencia, indica cómo efectuarías la búsqueda y los problemas que enfrentarías con:

      - Un diccionario con palabras desordenadas.
      - Un diccionario con palabras ordenadas, pero sin índice
      - Un diccionario con palabras ordenadas y con índice.

   j. Investiga por qué surgieron los sistemas NoSQL en la década de 2000 y compara a través de una tabla sus características vs. los sistemas de bases de datos tradicionales.

2. **Lectura**

   a. Leer el artículo Data's Credibility Problem y realizar un resumen del documento, destacando los puntos que a su consideración sean los más relevantes (no más de una cuartilla).

   b. Realizar un ensayo donde expreses tus comentarios (cada integrante del equipo deberá indicar este punto de forma individual en el documento que redacten) sobre la lectura, considerando los siguientes puntos:

      - Deberás indicar cuál es el objetivo que quiso plantear el autor: qué intenta decir, de qué intenta persuadirnos y/o convencernos, ¿cómo se relaciona con la materia de Fundamentos de Bases de Datos?
      - Deberán indicar cuál es la temática central del artículo y se deben señalar el tema o los temas laterales que desarrolla el mismo y cómo estos tienen relación con tú práctica profesional

- Consideraciones personales: deben indicar una postura ante las ideas planteadas en el artículo, proporcionar argumentos a favor o en contra (propios).

Respuestas

1. Conceptos generales

   a. Las principales ventajas son:

   - Acceso rápido a los datos.
   - Evita datos repetidos o duplicados.
   - Aumenta la productividad.
   - Permiten ingersar datos ilimitados.
   - Centralizar la información.
   - Reducción de espacio físico.
   - Mantenimiento fácil.
   - Permiten hacer respaldos.
   - Son portables.
   - Son dinámicas.

   Las principales desventajas son:

   - Pueden crecer mucho.
   - Sube costos.
   - Actualizaciones.
   - Pueden fallar críticamente.
   - Son presas de ataques remotos y virus.

   b. El esquema conceptual es aquel que define una descripción lógica básica, única y global que sirve de referencia para los otros esquemas. El esquema interno o físico es la descripción de cómo se guardan físicamente los datos en el sistema, esto incluye lo definido en el esquema conceptual, además de otros componentes como índices, tamaño de página, etc. Por último el esquema externo define los datos que cada usuario, según su rol en el sistema, puede y requiere ver. Por ejemplo un usuario de recursos humanos requiere ver los datos de cada empleado en la nómina de la empresa, mientras que un trabajador solo requiere poder manipular sus propios datos y datos que se involucren en sus tareas diarias.

   Cuando se hace un cambio sobre el esquema conceptual no debería afectar al esquema externo, es decir que los usuarios seguirán viendo los datos que requieren para realizar sus tareas a pesar de estos cambios. También si existen cambios en el esquema externo, como puede ser la modificación o creación de una vista, no afectarán a los usuarios que no estén involucrados en dicha modificación. A esto se le llama independencia lógica.

   Cuando hay cambios en el almacenamiento de los datos como cambiarlos de soporte de almacenamiento, cambiarlos de ubicación dentro del mismo soporte de almacenamiento o cambios en el acceso a registros determinados, formato o codificación no implica que hayan cambios en el esquema conceptual o que los usuarios en el esquema externo perciban alguna diferencia. A esto se le llama independencia física.

   c. La independencia de datos es una propiedad de los SMBD que permite que se realicen cambios en un nivel del esquema de datos sin tener que afectar a los otros niveles. Existen dos tipos de independencia de datos, física y lógica. Es más difícil obtener la independencia física ya que ahí radican los datos por lo que si se cambia el tipo de base de datos o un servidor definitivamente debe de haber un cambio en el código controlador de estos datos, rara vez es completamente agnóstico.

d. Aunque son términos que comúnmente se aceptan como sinónimos, existe una diferencia y radica principalmente en el momento en el que pueden comenzar a trabajar en el proceso de construcción de la base de datos.

El analista es quien decide, dadas las características de los datos con los que se trabajará y las necesidades de la empresa, que tipo de base de datos es la que mejor se acomoda. Plantea además que software y personal va a requerir el proyecto.

El Diseñador aterriza el proyecto en los esquemas conceptuales, divide y delega las tareas que mantendrán operativa la base de datos.

El Desarrollador programa y mantiene el código sobre el que trabaja la base de datos.

e. La base de datos es el conjunto de datos almacenado y ordenado de acuerdo a un modelo lógico. Por otro lado el SMBD es el programa que permite el acceso y la manipulación de los datos guardados.

Si bien el SMBD es la estructura sobre la cual podemos trabajar con los datos de manera eficiente, un buen diseño debe permitir poder cambiar de SMBD sin afectar la integridad de la base de datos.

f. El usuario ha tenido la oportunidad de conocer a grandes rasgos SMBD columnares, en memoria y aquellos que usa día a día son SMBD relacionales. Encuentra que en general el uso del lenguaje SQL es de utilidad, porque a pesar de que cada SMBD funcione internamente de forma diferente todos los manejadores que ha conocido utilizan SQL para manipular los datos, con alguna variación específica para cada SMBD pero en esencia la mayoría de las instrucciones son las mismas. Otra utilidad es que se puede separar la lógica de negocio del manejo de los datos, es decir que no se tiene que programar alguna funcionalidad para ordenar, consultar subconjuntos de información, etc. sino que todas estas tareas se le pueden dejar al SMBD y solo pedir lo que se requiera en el momento.

Las instalaciones que ha realizado son de MySQL, PostgreSQL y Sybase ASE, de las cuales casi no recuerda el proceso de instalación de Sybase, pero comenta que instalar MySQL y PostgreSQL en Linux es sencillo, ya que los sitios oficiales ponen los pasos para las distribuciones que ha manejado.

Como ventajas, todo SMBD se encarga de hacer consultas, agrupar y ordenar datos, hacer actualizaciones sin mayor complicación. Como desventaja cree que el uso directo de las sentencias SQL hace que hayan más probabilidades de realizar operaciones equivocadas, pero esto se arregla con el uso de un programa con una interfaz de usuario para dejar que el usuario pueda meter el menor número de datos posibles y mejor los seleccione de listas ya predefinidas. Otra desventaja es que a la larga existen estadísticas en los SMBD que si se van desactualizando harán que las consultas tarden mucho más, ya que no se tiene la información de cuántos datos hay en cierto rango de valores por ejemplo, lo que hace que un administrador de base de datos deba estar al tanto del rendimiento del SMBD.

g. Los sistemas NoSQL surgieron como respuesta al internet. Cuando los usuarios "normales" de internet empezaron a generar contenido era practicamente imposible diseñar un systema relacional confiable ya que los cambios que requeria eran muy constantes. Un esqema flexible permitio que se almacenara la informacion sin tener que adaptarla forzosamente a un esquema rigido.

independencia de datos	**Bases NoSQL**
Usan SQL para consultas de datos	La consulta de datos depende de la implementacion
Se basan en un modelo relacional	Tipicamente guardan pares llave-valor
Sistemas rigidos y controlados	Los esquemas son flexibles y adabtables
Los atributos en una tabla solo tienen un valor unico	Los documentos que guarda pueden tener atributos multivaluados

h. Es importante tener una base de datos para administrar un sitio web de manera eficiente, en especial uno donde se publican vídeos. Es importante evitar la inconsistencia y redundancia ya que podría haber múltiples usuarios con el mismo correo electrónico llevando a problemas al momento de asignar un dueño a un vídeo, una restricción en la base de datos hace manejar la redundancia trivial y evita la inconsistencia. Al estar todos los datos en una sola base de datos son accesibles para cualquier sistema de manera sencilla ya sea directamente a la base (un administrador) o

mediante una aplicación (usuarios). De esta manera se aísla el acceso a la información ya que solo el administrador ve la base completa mientras que los usuarios únicamente ven y manejan información respecto a sus cuentas. Las restricciones en la entrada de datos permiten que se eviten problemas de consistencia de esta manera un usuario no puede ingresar un valor invalido y solo es necesario aplicar la regla en la base de datos no en cada aplicación. Al ser us sistema transaccional las operaciones de SQL se realizan en su totalidad o no se realizan, esto garantiza que al usar una base de datos no haya problemas con transacciones parciales. Finalmente existe una garantía de que los valores son correctos incluso si se realizan cambios simultáneos justamente por que se manejan como transacciones y no realiza dos operaciones con exactamente el mismo valor.

i. Las principales responsabilidades de un Sistema Manejador de Bases de Datos son las siguientes:

- Abstracción de la información; Si no se cumple esta responsabilidad se corre el riesgo de guardar información innecesaria y que la Base de Datos empiece a crecer mucho debido a este almacenamiento extra que se tiene.
- Redundancia mínima; Si esta no se logra habría datos repetidos e impactaría en el tamaño de la Base de Datos así como en su desempeño en cuanto a búsquedas de información en esos campos. Así mismo elevaría la complejidad de extraer información y de darle mantenimiento.
- Consistencia; Si la Base de datos es inconsistente está podría generar que los datos no se actualicen de forma correcta, sobre todo en aquellas Bases de Datos en las que no se logra una Redundancia mínima, eso porque no se actualizan los datos en todas las tablas en donde aparecen.
- Seguridad; Si no se tiene una Seguridad en la Base de Datos esta puede tener problemas de consistencia de datos debido a usuarios malintencionados o frente a ataques que deseen extraer, manipular o destruir información de la Base de Datos o también a descuidos de algún usuario autorizado pero no competente para manipular la información de la Base de Datos.
- Integridad; De nada sirve una Base de Datos que no tenga datos íntegros y verídicos, esto debido a que las consultas que se realicen en esta Base de Datos no tendrá datos útiles.
- Respaldo y recuperación; En caso de un fallo catastrófico no se podría recuperar la información que estuviera en esa Base de Datos y se tendría que rehacer, así mismo para volver a un punto pasado, si no se tiene un respaldo este no podría llevarse a cabo.
- Control de concurrencia; En caso de no tener este control se podría dar que dos o más usuarios accedan al mismo dato y lo modifiquen, lo que causaría inconsistencias en la Base de Datos.

j. Para los 3 incisos se asume que el diccionario es un libro físico, el orden es alfabético y se esta buscando la palabra en una de las páginas.

- **Un diccionario con palabras desordenadas:** No habría una manera sencilla de buscar en el diccionario, debido a que no hay un orden la palabra se podría encontrar en cualquier página. La mejor manera de encontrar la palabra seria ir desde el principio buscar hasta encontrar la palabra. Esto es un problema porque es demasiado ineficiente, a menos de que se marque la página de alguna manera con palabras buscadas frecuentemente siempre que sea necesario hacer una búsqueda tendríamos que ir buscando desde el principio hasta encontrar cualquier palabra.
- **Un diccionario con palabras ordenadas, pero sin índice:** Al estar ordenadas las palabras entonces podemos hojear el libro para encontrar el principio y fin de la sección para la letra **L** como buscamos la palabra `"Luminiscencia"` debería de estar cerca del final entonces hojeamos en la sección desde atrás hasta encontrarla. Esto es mucho más eficiente ya que es fácil encontrar un área donde está la palabra de manera rápida pero aun así es necesario buscar en toda esta sección hasta encontrarla.
- **Un diccionario con palabras ordenadas y con índice:** Para buscar la palabra bastaría con ir directo al indice y ahí buscar `Luminiscencia` el indicé nos indica en donde se encuentra y podemos ir directo al resultado. Es la manera de búsqueda mas eficiente ya que no tenemos que buscar en una sección o en todo el diccionario solo en unas cuantas hojas donde está el indice (que asumimos está ordenado).

2. Lectura

   a. A continuación hay un pequeño resumen sobre la lectura *"Data's Credibility Problem"*

   Existe una infinidad de fuentes que producen datos para ser almacenados y usados/analizados después de un tiempo. Estos datos usualmente vienen de fuentes que son, en teoría, confiables por ejemplo un analista o un sensor especializado por lo que rara vez se revisa que los datos se encuentren "limpios" al momento de ser creados. Entonces cuando un usuario revisa los datos a detalle se da cuenta de que hay varios errores. Esto puede ser desde algo mínimo como un nombre sin acento en un documento hasta críticos como un error de medición en un instrumento medico.

   La responsabilidad de estos errores tiende a caer en quienes administran la base de datos, personas encargadas de TI. Esto es un error ya que a pesar de que ellos tienen la capacidad de corregirlo no siempre tienen la visión correcta del negocio o el uso que se le dará a los datos. Quien debe de ser el encargado de "limpiar" los datos es quien los usa. Esto trae un problema grande ya que limpiar datos es un proceso tardado, lento y costoso por lo que es mejor enfocarse en crear procesos que involucren a personas en distintos departamentos para que los datos sean limpios desde su concepción.

   b. Los ensayos de cada integrante se encuentran a partir de la siguiente pagina.

Ensayo Sela

Al leer este articulo inmediatamente me vino a la cabeza una anécdota que sucedió durante mi trabajo. Recuerdo claramente un problema que tuvimos con un cliente (una aerolínea que permanecerá sin ser nombrada debido a un acuerdo de confidencialidad activo). Dicha aerolínea necesitaba reportes de los usuarios que hablaban con el bot instalado en la pagina de facebook y cuantos lograban hacer una compra. La captura de los datos era mi responsabilidad ya que yo programé el bot. Obteníamos información básica como nombre, correo, teléfono y hasta que punto del flujo de compra llego el usuario. Pero al presentar los datos hubo muchas quejas debido a que al momento de contactar a varios usuarios los números telefónicos eran incorrectos. Resulta que asumíamos los teléfonos eran de la ciudad de México y se agregaba el código LADA de CDMX por omisión. Pocos usuarios daban un código LADA por lo que la mayoría de los datos eran incorrectos. Lejos de corregir los datos individualmente (que eran decenas de miles) se decidió implementar un sistema para determinar la ubicación del usuario y así asignar el código LADA correcto. los errores recudieron dramáticamente después de esto.

Justamente como vemos en el articulo es evidente que los encargados de los datos no siempre toman las decisiones correctas. En este caso asumir que los usuarios serian de CDMX fue un error que solo se notó hasta que los usuarios de los datos (los representantes telefónicos) trataron de usar estos datos y aunque trataron de corregirlo "al vuelo" no era posible por la cantidad tan grande. Enfocarnos en producir datos limpios mejoro la situación inmediatamente tras establecer comunicación con los usuarios de los datos.

Definitivamente concuerdo con los puntos que Redman establece en su articulo e incluso añadiría que debería ser parte del proceso standard de diseño integrar a los usuarios de los datos para saber exactamente como serán usados y así desde un principio poder generar datos limpios.

Ensayo sobre el artículo "Data's credibility problem" de Thomas C. Redman por Ernesto Cárdenas Torres.

Este artículo nos da una visión del porque es necesario contar con datos de buena calidad y nos presenta un panorama del flujo de datos para sugerir acciones a tomar en caso de detectar anomalías en los mismos.

Por ejemplo, en el caso de la Petrolera Chevron, donde se detectaron datos incompletos con un tiempo de reparación que sobrepasaba los 5 años, se sugiere enfocar esfuerzos en asegurar la calidad de los datos nuevos que se capturen, para lograrlo se emplea un cambio en las métricas con las que se evaluaba la confiabilidad de un documento.

Además es mencionado que la comunicación es clave. Tal es el caso entre los trabajadores de la compañía, si uno detecta datos erróneos debe comunicarlo al resto, no solo corregirlos, de esta manera pueden verificar conjuntamente si se trata de un error aislado o si es todo el conjunto de datos el que está dañado. Adicionalmente debe existir comunicación entre aquellos que proveen los datos y aquellos que los usan, de esta manera las metas se logran de mejor manera y ambas partes trabajan más eficientemente.

Por otro lado, es mencionado que se suele asociar al departamento de TI con las problemáticas de los datos, y es que en primera instancia es el área encargada de resguardarlos, sin embargo, como parte activa de esa área, si tenemos una idea de como es que están circulando los datos y que hace la empresa para mantener la calidad de los mismos podemos ayudar a detectar áreas de oportunidad para agilizar la recuperación de la credibilidad.

Finalmente el artículo explica brevemente la importancia de verificar los metadatos, ya que en la actualidad llegan a ser tan útiles como los datos mismos.

Concuerdo con las ideas que nos presenta el autor. Uno, la importancia de contar con datos creíbles es innegable. Dos, la comunicación; las empresas deben aceptar cuando los datos son de mala calidad e instruir al personal a que sea parte activa en la verificación de los datos al momento en que trabajen con ellos y así reportar cualquier anomalía con prontitud. Tres, es necesario que se desmitifique la idea de que la calidad de los datos depende exclusivamente del área de TI.

Ensayo Dadmy Nolasco

El problema de la calidad de datos es un tema del cuál no muchos nos preocupamos, o al menos no a fondo. Sin embargo hoy en día es el activo más importante de cualquier empresa después de sus empleados.

A lo largo de la lectura se puede observar que el autor saca a relucir las prácticas no favorables que se dan con los datos. A menudo nos encontramos con un problema que tiene que ser arreglado lo más pronto posible, como es el caso de la ejecutiva cuyos datos en el reporte fueron arreglados antes de su presentación, pero una vez arreglado esto olvidamos la parte esencial que es verificar cuál fue la falla o en dónde se corrompieron estos datos, y simplemente continuamos con nuestro trabajo hasta que se presenta otro incidente.

Comparto la opinión del autor de que la calidad de los datos debe cuidarse desde su nacimiento. En nuestro curso de fundamento de bases de datos estamos aprendiendo que un buen diseño puede evitar problemas posteriores, pero también es necesario que haya algún estándar para manejarlos, ya que si no puede ocurrir un problema como el de la NASA donde se utilizaron dos sistemas métricos diferentes en una operación. También he sabido de casos donde un dato como el género lo han representado de diferentes formas en distintos departamentos de una misma empresa, con valores como "M" y "F", "H" y "M" o "1" y "2", por lo que no hay consistencia.

Una parte interesante que comenta el autor es que si se corrige la calidad desde el origen de los datos se pueden ahorrar muchos recursos. No obstante también plantea que departamentos como IT no se sienten incentivados para mantener todo en orden, además de hacer falta comunicación entre aquellos que ven nacer los datos y aquellos que los utilizan.

En lo personal creo que en muchas empresas es difícil que todos cuiden la calidad de datos incluso si les indican que es de suma importancia para cierta operación, pero también creo que poco a poco se puede mejorar estableciendo estándares para que todos estén en el mismo canal.

Ensayo Yamil

El artículo "Data's Credibility Problem" tiene como objetivo explicar la importancia de la calidad de datos, cuales son los problemas con la calidad de los datos, por ende, como surgen estos problemas y de que forma se intentan corregir, así como una recomendación sobre lo que se tiene que realizar para poder tener una mejor calidad de datos. El problema de la calidad de datos se centra muchas veces en que el DBA no está familiarizado con los datos que este manejará en la BD y por ende muchas veces estos datos no se manejan de la forma correcta o no se estructura la BD de la forma en la que estos datos realmente cobren sentido. También se tiene el tema de la calidad de datos cuando el origen de los datos es defectuoso. Al final de cuentas los usuarios finales son los que sufren por esta mala calidad de datos, ya que estos son los que entregan cuentas y/o presentan datos en reuniones, datos que no están correctos. Si el tema de la calidad de datos en una organización no se resolviera desde la creación, recomendaría usar un proceso ETL (Extract, transform, Load) en donde se tenga un paso de validación que homologue y/o valide los datos para asegurar una calidad de datos superior a la que se tenía. El tema de calidad de datos siempre es un tema importante pero que muchas veces las organizaciones dejan de lado, en mi experiencia con este tema, muchas veces las empresas no tienen noción de la importancia de la calidad de datos en sus BD hasta que alguien externo llega y empieza a realizar consultas que al final arrojan datos y/o resultados que no hacen sentido al área que pidió esa consulta. Una vez llegado a este punto, el tema de la limpieza y la mejora de la calidad de datos es un tema que se vuelve prioritario pero al cual no se introducen de forma

necesaria, es un tema en el que saben que se realiza, pero realmente no les interesa saber como se realiza, para la empresa u organización es solo importante que se de y para esto es importante entender como se crean los datos y como se alimentan al BD, ya que en ocasiones el problema no está en la creación de datos si no, en la carga de los datos a la BD; En ocasiones muy raras el problema se debe a la estructura incorrecta de la BD. Por lo anterior, mi percepción sobre el tema se mantiene ya que son casos que me ha tocado experimentar en carne propia.

Ensayo Jose Angel