Tarea 1 Fundamentos de Bases de Datos

Ayala Morales Mauricio
ContrarioMotu@ciencias.unam.mx

Díaz Tinoco Gisel Maite
digit@ciencias.unam.mx

Gómez Vázquez Carlos Alberto
carlosGomezMX@ciencias.unam.mx

Gutiérrez Medina Sebastián Alejandro
sebasguti1511@ciencias.unam.mx

Ramírez Gutiérrez Oscar
rg.oscar17@ciencias.unam.mx

11 de marzo de 2022

1. Conceptos generales:

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

Ventajas:

- Permite crear una realidad unificada con la cual generar conocimiento.
- Minimaliza la redundancia que puede haber en los datos.
- Garantiza la consistencia de los datos.

Desventajas:

- El proceso de creación de una base de datos es largo.
- Se debe dar un mantenimiento o dar solución a posibles fallas.
- Necesitad de capacitación al personal.

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

La independencia de datos es una propiedad de los SMBD, permite cambiar el esquema de un nivel a otro sin tener que cambiar el esquema de la base de datos. Sirve para mantener la información separada de todos los programas que la utilizan. Existe la independencia física y la lógica.

La independencia de datos física ayuda a separar los niveles conceptuales de los internos o físicos. Es decir, ayuda a dar una descripción lógica de una base de datos sin la necesidad de meterse con el aspecto físico. Con independencia física se pueden cambiar las estructuras de almacenamiento sin que esto represente un cambio interno dentro de la estructura.

La independencia lógica se da cuando se pueden hacer cambios en el esquema conceptual sin afectar a las vistas externas o los programas.

Es más difícil el lograr la independencia lógica ya que tiene que haber un muy buen análisis de requerimientos para realmente tener una buena representación de los datos antes de empezar a construir todo el sistema.

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?

En los esquemas **externos**, se sitúan las diferentes vistas lógicas que los procesos usuarios tendrán de la base de datos que utilizaran.

En los esquemas **conceptuales**, es una única descripción lógica, básica y global que sirve de referencia para los demás esquemas.

En los esquemas **físicos/internos**, es una sola descripción física.

La **independencia de datos lógica** se garantiza si logramos separar los esquemas externos de los esquemas conceptuales. Mientras que la **independencia de datos física** se garantiza si logramos separar los esquemas conceptuales de los esquemas físicos/internos.

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.

Los analistas de bases de datos se encargan de recopilar la información de manera cualitativa, que se da mediante el diálogo con el experto de su propia empresa, y la cuantitativa que se obtiene mediante los datos numéricos. El propósito de recabar estos datos es el examinar e interpretarlos para así poder tomar decisiones empresariales en base a ellos.

El diseñador de bases de datos se encarga de los detalles del diseño específicos de la base de datos, como las tablas, vistas, procedimientos, etc. Este diseño debe de concordar con la información que le fue proporcionada por el analista de bases de datos para asegurarse de que se estén teniendo en cuenta las necesidades de la empresa en cuanto al diseño de la base de datos.

El desarrollador se encarga de la seguridad e integridad de la base de datos, además que dependiendo, pueden encargarse de la conexión de la base de datos a un sistema existente o desarrollar una interfaz de usuario.

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

Una base de datos es una colección de datos relacionados con objetivo en común. Un SMBD es un software que nos facilita el procesos de definir, construir, manipular y compartir datos en una BD para las diversas aplicaciones. Por lo que una BD es administrada por un SMBD.

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?

Entrevistamos al ayudante de la materia de Computación Distribuida:

"Encuentro útil que pueden controlarse desde interfaces (en el caso de java usando JPA) pues te olvidas en sí del SMBD"

No hubo una inclinación hacia una específica sin embargo se nos dijo que "lo que requiere más esfuerzo es hacer la conexión con la base de datos porque luego instalar las dependencias y manejarlas es muy talachudo"

"La principal ventaja que los usuarios ven es que tiene una manera más eficaz de acceder a información. Siendo muy común que tengan toda su información en excel. Aparte con un buen SMBD pueden tener operaciones descentralizadas, lo cual ayuda a que no concentren su información en un sólo excel. El problema es que en

verdad en todos lados no hay gente que sepa de bases y es por eso que muchas veces se deciden por excel. Y también que sus datos no son tan grandes como para ameritar una base"

g) Supón que deseas crear un sitio de vídeos 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 vídeos: título, el usuario que lo subió, la fecha de carga, las etiquetas, qué usuarios lo vieron, cantidad de "Me gusta", entre otros.

- **Redundancia de datos e inconsistencia** : Ya que un vídeo puede tener más de una etiqueta, el mismo vídeo puede estar almacenado en varios archivos que consistan de cada una de las diferentes etiquetas existentes, esto generaría un gran gasto de almacenamiento

- **Dificultad al acceder a los datos** : Usando el ejemplo anterior, si se quiere buscar a qué etiquetas pertenece un vídeo, se tendría qué hacer una búsqueda por cada uno de los diferentes archivos de cada una de las etiquetas, decidiendo si ese encuentra en el archivo o no. Esto sería completamente ineficiente para algo tan simple como obtener la información de un solo vídeo.

- **Aislamiento de datos** : Ya que los vídeos están almacenados en diferentes archivos puede que existan archivos con formatos diferentes, entonces si se quisiera hacer la búsqueda de, por ejemplo, todos los vídeos de un cierto usuario, o los vídeos que se subieron en cierta fecha, podrían ocurrir fallos por las diferencias de cómo se hizo cada archivo.

- **Problemas de integridad** : Cualquier modificación que un usuario quiera hacerle a sus vídeos, como editarlo, editar su nombre o descripción, el programa que tenga qué realizar estas operaciones tendrá qué modificar el mismo vídeo en cada archivo en el que se encuentre (por la redundancia de datos), esto haría que añadir modificaciones en un futuro sea una tarea tediosa y abierta a errores, pues podría ser mal implementada y no modificar todas las copias del mismo vídeo.

- **Problemas de atomicidad** : También, al realizar cualquier acción o modificación, como dar "Me gusta.º cambiar el título, podría suceder un fallo en el sistema y que alguna de las copias del vídeo en alguno de los archivos no se actualice correctamente.

- **Problemas de acceso concurrente** : Si, por ejemplo, dos personas le dieran "Me gusta.ªl mismo vídeo exactamente al mismo tiempo, la cantidad total de "Me gusta"podría quedar incorrecta, pues cada proceso estaría tomando la misma cantidad y la modificaría solamente sumándole 1, por lo que el nuevo total quedaría como si solamente una persona le hubiera dado "Me gusta". Esto puede suceder de la misma forma si dos personas empiezan a ver el mismo vídeo al mismo tiempo, la cantidad de vistas quedaría incorrecta.

- **Problemas de seguridad** : Si un usuario que sube vídeos decide poner uno de ellos privado, por los anteriores problemas de integridad y redundancia, cualquier persona podría encontrar ese mismo vídeo en otra etiqueta (archivo), pues puede que no se haya actualizado correctamente la privacidad del vídeo, esto le permitiría a cualquier persona tener acceso a cualquier vídeo privado.

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.

Definición : Involucra especificar los tipos, estructuras y restricciones de los datos almacenados. Si no especificamos los tipos de nuestros datos, podríamos agregar cualquier dato y lo tomaría como correcto.

Construcción : Proceso de almacenar los datos en algún medio de almacenamiento controlado. Si lo almacenáramos en un lugar donde no tengamos control no podríamos garantizar la integridad y acceso de los datos.

Manipulación : Tener funcionalidades tales como consultas, recuperación y actualización de datos. Si no podemos modificar los datos, tendríamos datos demasiado estáticos, lo cual no es muy útil.

Distribución : Permite el acceso a más de un usuario. No podríamos crear una realidad unificada, solo tenemos nuestra realidad.

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.

  Dado que las palabras están desordenadas, tendríamos que buscar palabra por palabra hasta encontrar Luminiscencia. Es decir una búsqueda lineal.

- Un diccionario con palabras ordenadas, pero sin índice

  Dado que las palabras están ordenadas, abriríamos el diccionario en cualquier página y ver si la letra con la que inicia es l, si este es el caso solo queda buscar en esa hoja. Si no vemos si la letra de donde abrimos nuestro diccionario es mayor que l o menor (tomando como el orden el alfabético), si es mayor entonces nos vamos por el lado izquierdo y si es menor nos vamos por el lado derecho y repetimos el proceso. Es decir una búsqueda binaria.

- Un diccionario con palabras ordenadas y con índice.

  Buscamos la palabra en el índice, obtenemos el número de la página en el que está la palabra y vamos a la página.

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.

El acrónimo NoSQL corresponde a la frase "Not only SQL", aunque originalmente era "No SQL". Este término hace referencia a cualquier sistema de gestión de datos que no sigue el modelo de las SMBD y que no aplica el lenguaje SQL para realizar las consultas.

Este tipo de bases de datos comenzaron a originarse en el año 1998 y no fueron creadas por una empresa específica, sino que fueron concebidas por distintas empresas y grupos independientes que buscaban soluciones específicas a sus problemas.

SQL	NoSQL
Relacional.	No relacional.
Basada en tablas.	Basada en documentos, pares de llave/valor o gráficas.
Mantienen una estructura definida.	Su estructura puede ser dinámica (en el caso de que sean orientadas a documentos).
Mantienen una mínima redundancia.	Permiten redundancia de datos.
Mantienen la consistencia en los datos, sin dar lugar a errores.	Mejora el rendimiento reduciendo la consistencia si es necesario, lo que permite alojar una gran cantidad de datos.
Escalables verticalmente, aumentando los recursos del servidor.	Escalables horizontalmente, aumentando la cantidad de servidores.

2. Lectura

1. 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).

   El artículo nos habla de cómo la calidad de los datos y la falta de comunicación entre departamentos puede terminar en pérdidas para la misma. Como nos da un ejemplo, se corrigen errores en un área pero no se le notifica a otra, por lo tanto, se sigue trabajando con datos erróneos lo cual sigue llevando a más errores futuros.

   Sin embargo, algo que dicen es que los trabajadores gastan alrededor de un 50 % de su tiempo corrigiendo errores en los datos, cosa que se podría solucionar o por lo menos disminuir el índice de error si en vez de corregir los errores sobre la marcha que además de ser un trabajo tedioso muchas veces también se deja a la interpretación de quien esté rectificando los datos, lo cual pues es malo por esa inadecuada manipulación de datos, sino que puede solucionarse cuando se crea nueva información. Esto quiere decir que el problema no debería de ser el revisar cada una de la información ya existente, si no que se debe de crear una comunicación desde el inicio para poder evitar errores desde la creación de esa información, al igual que no mantener a los trabajadores en blanco ya que el comprender el porqué es necesario que la información que se está recaudando es importante que sea fiable es más fácil que el trabajo se haga bien si comprenden el porqué se debe de ser preciso.

2. 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:

   a) 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?

   b) 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

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

   - **Oscar Ramírez Gutiérrez**

Lo que el autor nos trata de decir es que mejorar la calidad de los datos es algo sencillo, que no tiene que resolver los especialistas en datos, sino ser resuelto entre los creadores y usuarios de los datos, con una buena comunicación entre ellos. Pienso que se relaciona con la materia, ya que como desarrolladores no debemos de olvidar que al definir una base de datos, tenemos que satisfacer las necesidades/requerimientos del cliente, en otro caso estaríamos haciendo algo que no le va a ser de utilidad.

Considero que la temática central del artículo serían qué papel toma la calidad de los datos en la mayoría de las empresas, cómo es que afecta en la productividad, toma de decisiones, etc. Cómo es que al potencialmente corregir localmente un error en los datos, al no comunicar de dicho error, este se va propagando, o peor al no tener confianza en los datos, las decisiones que se tomen no serán con base en los resultados de analizar los datos. Concluye que la mayoría de las veces no se tiene esa calidad en los datos, porque los directivos no le toman suficiente importancia o no saben como mejorar su calidad.

Algo que enfatiza el autor y estoy de acuerdo, es la importancia de la comunicación entre los desarrolladores, los que generan los datos y los que usan esos datos. Además de que se haga entender que al que genera los datos que alguien va a utilizar sus datos para tomar decisiones.

- **Gisel Maite Díaz Tinoco**

El objetivo del autor era el hacernos saber el como el manejo de la información puede optimizarse de tal manera que se reduzcan los costos dentro de la empresa ya que la mayor parte del tiempo de trabajo en ella es la corrección de datos y mientras que a fin de cuentas es algo necesario, podría ahorrarse tiempo si la información estuviera bien capturada desde la creación de esos datos. El mantenimiento aún así va a requerir la captura de errores dentro de estos datos, pero al menos podría decirse que el porcentaje de error va a ser menor, lo cual reduciría el costo de los errores que se lleguen a cometer a partir de esa información errónea como en el costo de trabajo de estar cachando errores.

Por otro lado, nos dice el cómo aún si se intenta capturar una información correcta desde el inicio, hay ocasiones en las que puede "mancharse" o "ensuciarse" debido a factores que en ocasiones no pueden ser controlados, como decían en el caso del petróleo, la persona encargada de las mediciones estaba haciendo su trabajo correctamente, sin embargo, existía un sesgo en la información debido a que la herramienta detectaba un nivel más alto debido a, literalmente, la suciedad hacía que se marcara un nivel más alto del que se suponía debería marcar. El artículo también recalca la importancia de la comunicación ya que este problema se resolvió comunicando el problema a los encargados de esto, pero igual se le informó el porqué ocurría el error al empleado que tomaba las medidas, de esta manera, es más fácil que el encargado de las mediciones pueda detectar cuándo se podría tener un error y podría dar aviso de que necesita limpieza. De la misma manera, con el primer ejemplo, se hizo una detección de error en los datos y se corrigieron en esa instancia, pero al no dar aviso a todos los departamentos (o al menos a la mayoría a la que le competa ese manejo de información), realmente la corrección no sirvió de nada ya que la empresa va a seguir operando con datos erróneos, lo cual no sólo hace que el problema continúe si no que si esta información es usada para otros procesos, va a causar un problema más grande en algún punto, y en algunos casos puede derivar en una pérdida monetaria.

Comprendiendo el error y comunicándose hace que sea más fácil una futura detección y además que haya varias personas que lo comuniquen y puedan reconocer el error.

- **Mauricio Ayala Morales**

El escritor describe la importancia de hacer correcta la recolección y el mantenimiento de datos, y los problemas que generaría hacerlo incorrectamente; de igual forma proporciona algunas técnicas que son utilizadas en la industria para corregir y administrar los montones de datos que ya han sido recolectados.

La primer medida propuesta es que desde el principio exista una buena comunicación entre los encargados de la recolección de datos, al igual que las personas que los utilizaran, lo cual es de suma importancia ya que si desde el principio todos los participantes tienen en claro cuál es el objetivo del proyecto de la

empresa, para así seleccionar los datos relevantes, provocando que la recolección de datos esté siendo concientizada y se desechen los datos que no son necesarios para cumplir los objetivos. Esto ahorraría gastos que se generarían al corregir los datos imprecisos, para no tener un cúmulo de datos inútiles de la empresa.

La segunda medida está dada a partir de la experiencia de una experta que se encontraba a cargo del equipo de gestión de datos, donde los casos donde existe una gran colección de datos es preferible perfeccionar la manera en que los nuevos datos sean recolectados que intentar reparar los ya existentes, lo cual es una técnica muy funcional para detener el problema, pero si los antiguos deben utilizarse se deberá realizar un esfuerzo por corregirlos.

Ambas técnicas pueden ser empleadas para diseñar las bases de datos que creemos en un futuro, puesto que si desde el comienzo se tiene en claro cuales son los datos relevantes para resolver problemas del proyecto el diseño será en base a la necesidad, provocando así que se tenga una integridad en los datos y una mejoría en la creación de sus bases.

- **Sebastián Alejandro Gutiérrez Medina**

El autor del articulo tiene como objetivo mostrarnos que al lidiar con problemas de datos incorrectos o desorganizados lo ideal no es corregir los datos actuales y anteriores sino identificar que esta generando el problema y solucionarlo para que los datos generados en el futuro no tengan el mismo problema, pues si solo se arreglan los que se necesitan en ese momento, el problema seguirá y se tendrá que repetir el proceso una y otra vez.

A su vez, debe existir una mejor comunicación entre los creadores de datos y los que utilizan dichos datos, pues si los creadores de datos supieran que uso se le dará a esos datos y/o cuales necesitan ser mas precisos que otros esto reduciría y prevendría errores en los datos, además de que al haber una comunicación mas directa ayudaría en crear e implementar nuevas maneras de mejorar la calidad de los datos e incluso incrementar la eficiencia con la que se procesan los datos.

Otro aspecto a considerar es que se debe de dejar de centralizar las responsabilidades del manejo y distribución de datos a un solo sector de una empresa, es necesario que si se llegan a implementar las anteriores ideas no solo sea en una parte especifica de la cadena de trabajo, si no sobre todas las partes involucradas, aunque puede que no sea posible una conexión completa entre todas y cada una de las partes, es importante que tampoco estén solo conectadas con la parte inmediatamente anterior y siguiente parte del proceso, pues cuando suceda un problema al final de todo y la solución involucre a una parte que no sea la inmediatamente anterior, se tendría que ir por cada uno de las partes de la cadena de trabajo hasta que se halle la raíz del problema lo cual es extremadamente ineficiente, cuando esto puede ser prevenido con una mejor comunicación con todas las partes del proceso y no solo con las adyacentes.

En conclusión es importante que al encontrarnos con datos erróneos o imprecisos analizamos y solucionemos primero la causa del problema y además de tratar de tener una mejor comunicación con las partes involucradas durante todo el proceso por el que pasaron los datos antes de llegar a nosotros.