Profesor: Gerardo Avilés Rosas
Ayudantes: Rocío Aylin Huerta González
Rodrigo Alejandro Sánchez Morales

1. Conceptos generales:

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

Ventajas	**Desventajas**
Redundancia de datos reducida.	El enfoque de datos es costoso debido a los mayores requisitos de Hardware y Software.
Mayor integridad de datos e independencia de los programas de aplicaciones.	Los sistemas de bases de datos son complejos (debido a la independencia de los datos), difíciles y lentos de diseñar.
Costos reducidos de entrada, almacenamiento y recuperación de datos.	Los daños a la base de datos afectan prácticamente a todos los programas de aplicaciones.
Reducción de errores de actualización y mayor consistencia.	Altos costos de conversión al pasar de un sistema basado en archivos a un sistema de base de datos.
Mejora de la seguridad de los datos.	Capacitación necesaria para todos los usuarios.

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

Es la capacidad de modificar el esquema sin afectar los programas y la aplicación a reescribir. Los datos están separados de los programas, por lo que los cambios realizados en los datos no afectarán la ejecución del programa y la aplicación, por lo cual decimos que los niveles superiores no se ven afectados por los cambios en los niveles inferiores.

Independencia de datos lógicos significa cambiar el esquema conceptual sin cambiar la API externa, los programas o las vistas externas. Mientras que la independencia de datos físicos es cuando el esquema de software se cambia sin necesidad de actualizar los programas de software para su modificación. Debido a lo expuesto anteriormente concluimos que es mucho más difícil lograr la independencia de los datos lógicos en comparación con la independencia de los datos físicos.

   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?

Externo:

- Permiten personalizar (y autorizar) el acceso a los datos a nivel de usuarios individuales o grupos de usuarios.
- Proporcionan independencia de datos lógicos.

Interno:

- Visto únicamente por el administrador.
- Describe cómo se almacenarán físicamente los datos y cómo se accederá a ellos.
- Permite que los datos almacenados en la base de datos se puedan recuperar mediante una sola operación.

Conceptual:

- Este es visto por el arquitecto y el administrador.
- Describe los datos almacenados en el modelo de datos.
- Define toda la base de datos sin hacer referencia a cómo se almacenan los datos en la memoria secundaria de la computadora.

Estos 3 niveles se relacionan ya que en el esquema conceptual se describen todas las relaciones que se almacenan en la base de datos, cualquier base de datos tienen exactamente un esquema conceptual y uno interno porque solo tiene un conjunto de relaciones almacenadas, pero puede tener varios esquemas externos cada uno adaptado a un grupo particular de usuarios. Los esquemas externos proporcionan independencia de datos lógicos, mientras que los esquemas conceptuales ofrecen independencia de datos físicos.

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.

El analista tiene que garantizar la utilidad de toda la información y los programas relacionados con los datos, incluido el sistema de administración de la base de datos, cualquier sistema de visualización de datos, la eficiencia del lenguaje de consulta de datos y el diccionario de datos.

El diseñador crea una estructura de base de datos para hacer frente a las necesidades y expectativas de los futuros usuarios, también debe desarrollar formas de mostrar la información a los usuarios, además de mantener y adaptar bases de datos existentes siguiendo las necesidades cambiantes de los usuarios, o las cambiantes posibilidades en la programación.

Los desarrolladores implementan el código de la base de datos para realizar una variedad de tareas, que incluyen: Extracción de datos de una base de datos para análisis de informes y crean, actualizan, extraen o eliminan datos según lo que requiera una aplicación e incluso diseñan nuevas bases de datos que satisfagan las necesidades de los usuarios o clientes, en un formato eficiente y preciso.

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

Apartir de la base de datos, el Sistema Manejador de Bases de Datos ó SMBD por sus siglas permite gestionar cómo se organiza y optimiza la información, es decir, es una **interfaz** entre el usuario y la base de datos. En esta, debe haber coherencia entre todos los datos de la base de datos y corrección y exactitud de la información contenida en la misma.
Los SMBD soportan un modelo de datos **relacional**, es decir, consiste en en una colección de relaciones, que contienen atributos de un tipo específico, lo cual en una base de datos puede modelarse mediante el **modelo entidad-relación**
Por esto mismo, un SMBD permite una supervisión y control mucho más **sencilla** de las bases de datos, ya que se pueden representar relaciones complejas entre datos y otros aspectos relacionados con la seguridad y validez de los 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?

Respuesta

- **¿qué características de SMBD encuentran más útiles y por qué?**

  - **La minimización de la redundancia/duplicación de datos** : Ya que esto permite que la información se mantiene concisa y solo aparece una vez para evitar la imprevisibilidad de los datos

  - **Se mantiene la integridad de los datos:** Esta caracteristca permite evitar la corrupción y falla de la base de datos y asegura que se mantenga la integridad de los datos.

- **¿qué instalación(es) de SMBD encuentran más/menos complicada y por qué?**
  (a) **Más fácil** PostgreSQL, nunca tuve problemas en instalarla tanto en Linux como Windows, es fácil ya que hay bastante documentación y videos sobre como hacer la instalación.
  (b) **Más complicada** MySQL, tanto en Linux como en Windows, me salía un error de que no podía conectar con la base de datos, en general los errores más pesados y más comunes son de ese tipo, los cuales a veces se deben a los requerimientos del sistema.
  Un ejemplo de estos errores es: **ERROR 2002 (HY000): Can't connect to local MySQL server through socket '/var/run/mysqld/mysqld.sock' (2)**

- **¿cuáles perciben estos usuarios que son las ventajas y desventajas de un SMBD?:**
  (a) **Ventajas**
    - La seguridad de los datos.
    - La inconsistencia de datos se minimiza.
    - El acceso a los datos.
  (b) **Desventajas**
    - La commplejidad de la gestión de un SMBD.
    - Los costos de manterner trabajando y actualizando 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.

Respuesta

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.

(a) **Define la base de datos-** Especifica los tipos y estructura que tendrán los datos. Si esto no se cumpliera, habría un problema de modelación en la base de datos, y por lo tanto no habría relación ni correctez.

(b) **Construir una base de datos:**-Por medio de un algoritmo, dados algún tipo de datos, los guarda en un medio controlado. Al igual que el caso anterior, pueden ocurrir problemas de relación entre los elementos de la base de datos en caso de que no se construya de forma correcta.

(c) **Manipular la base de datos-** Actualizar la base de datos, realizar consultas. Si ocurre un problema en la actualización de la base de datos, puede que datos se pierdan, es decir, no será **consistente** ni **Íntegra**

(d) **Seguridad**-Debe garantizar que se restringe la cantidad de información a los usuarios. Si esta responsabilidad fallara, la base de datos quedaría expuesta, pudiendo filtrar información priviliegiada a personas malintencionadas

(e) **Respaldo y recuperación-** Estas deben proporcionar una forma eficiente de realizar copias de respaldo de la información almacenada en ellos, asegurando que se pueden restaurar datos a partir de estas copias. Si no se generan copias, se puede perder información muy valiosa que será muy costoso volver a modelar si es una base de datos grande.

(f) **Control de la concurrencia-** Lo más habitual es que sean muchas las personas que acceden a una base de datos, bien para recuperar información, bien para almacenarla. Si esto no se hace de forma correcta, pueden ocurrir inconsistencias en nuestra base de datos.

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:

(a) Un diccionario con palabras desordenadas.

(b) Un diccionario con palabras ordenadas, pero sin índice.

(c) Un diccionario con palabras ordenadas y con índice.

a) *Un diccionario con palabras desordenadas-* Al querer buscar la palabra en este diccionario, no habría una forma específica de encontrarla. Lo único que podríamos hacer, es ir leyendo todas las palabras en todas las hojas del diccionario hasta dar con la palabra. El problema de este diccionario es que al haber tantas palabras en un diccionario, la busqueda será un proceso largo y muy tardado, y a la mínima falta de concentración podemos pasar de largo la palabra y nunca encontrarla.

b) *Un diccionario con palabras ordenadas, pero sin índices-* Para este diccionario, podríamos recorrer algunas hojas de forma **arbitraria** para irnos acercando a las palabras con letra L. Cuando lleguemos a este índice, empezamos a revisar ahora las palabras que empiecen con Lu, y así hasta llegar a la palabra completa.

c) *Un diccionario con palabras ordenadas e índices-* Para este diccionario la búsqueda es muy sencilla, sólo buscamos el índice L, y después buscamos las palabras que empiecen con Lu, de esta forma encontraremos la palabra de forma rápida. Esta es la forma más cómoda y sencilla de encontrar una palabra en el diccionario.

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.

Gigantes de Internet como Facebook, Google, Amazon vieron aumentos repentinos en el tráfico y los datos. Las bases de datos tradicionales no podían escalar bien, además de que esto es una tarea costosa y que requiere mucho tiempo. Los desarrolladores eran el costo para la empresa en lugar del almacenamiento.

Profesor: Gerardo Avilés Rosas
Ayudantes: Rocío Aylin Huerta González
Rodrigo Alejandro Sánchez Morales

A fines de la década de 2000, surgieron las bases de datos NoSQL donde se cambiaron los modelos de datos complejos y difíciles de administrar para evitar la duplicación de datos. Las bases de datos NoSQL lograron disminuir los costos de almacenamiento y los costos de los desarrolladores. Optimizando las actividades diarias de gestión de datos.

Debido a la caída de los precios del almacenamiento de datos, aumentó la demanda de almacenamiento y consulta. Los datos venían desde datos estructurados y semiestructurados hasta datos polimórficos. Ahí es donde las bases de datos NoSQL manejaron todas las demandas de almacenamiento de datos no estructurados. Ayudando así a los desarrolladores a almacenar datos y proporcionándoles una mayor flexibilidad.

Bases de datos Tradicionales	**NoSQL**
Las bases de datos tradicionales están basadas en tablas en forma de filas y columnas y deben adherirse estrictamente a las definiciones de esquema estándar.	Las bases de datos NoSQL pueden basarse en documentos, pares clave-valor, gráficos o columnas y no tienen que ceñirse a definiciones de esquema estándar.
Lenguaje de consulta estructurado	Tienen el esquema dinámico para datos no estructurados. Los datos se pueden almacenar de forma flexible sin tener una estructura predefinida.
Costos reducidos de entrada, almacenamiento y recuperación de datos.	Los daños a la base de datos afectan prácticamente a todos los programas de aplicaciones.
Costoso de escalar.	Las bases de datos NoSQL son escalables horizontalmente. Esto significa que al fragmentar o agregar varios servidores a esta base de datos, puede manejar un mayor tráfico.
Agregar nuevos datos en la base de datos tradicionales requiere que se realicen algunos cambios, como el relleno de datos, la alteración de esquemas.	Más barato de escalar en comparación con las bases de datos tradicionales.
	Los nuevos datos se pueden insertar fácilmente en las bases de datos NoSQL, ya que no requiere ningún paso previo.

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).

Los problemas de calidad de datos es un tópico que afectan a todos los departamentos, industrias, niveles y tipos de información.

Los estudios muestran que los trabajadores de conocimiento pierden hasta el 50% del tiempo buscando datos, identificando, corrigiendo errores y buscando fuentes de confirmación para datos en los que no confían. Todo esto nos lleva al lema "basura que entra, basura que sale".

Existen 2 momentos importantes en la vida de los datos:

1. El momento en que se crean.

2. El momento en que se usan.

La calidad de datos se fija en el momento de la creación, pero en realidad no juzgamos esa calidad hasta el momento de su uso.

La solución no es una mejor tecnología: es establecer una mejor comunicación entre los creadores de datos y los usuarios "clientes" de datos, asegurando que estos sepan cómo se usaran los datos para que así puedan identificar las causas fundamentales de los errores y encontrar formas de mejorar la calidad en el futuro.

En lugar de un esfuerzo masivo para limpiar los datos incorrectos existentes las empresas deberán centrarse en mejorar la forma en que se crean los nuevos datos. Ya que la forma en que se crean nuevos datos, se deben identificar y eliminar las causas fundamentales del error. Una vez hecho eso, se requiere de una limpieza limitada, pero no una limpieza continua. Con la observación de que pongan la responsabilidad de los datos en manos de los gerentes en línea ya que los creadores de datos no están vinculados organizacionalmente a los usuarios de datos.

Muchos de estos problemas de calidad de datos se dan en los metadatos "datos sobre datos". Los metadatos de alta calidad facilitan que las personas encuentren los datos que necesitan, combinen información y saquen las conclusiones apropiadas, por otro lado los errores en los metadatos pueden tener un gran impacto.

Las barreras reales para mejorar la calidad de datos son algunos gerentes que se niegan a admitir que sus datos no son lo suficientemente buenos y otros simplemente no saben cómo arreglar los datos de mala calidad.

Sin duda el primer avance ocurre cuando un gerente en algún lugar de la organización (posible alto ejecutivo), se cansa y decide iniciar un programa de datos para así mejorarlos.

En conclusión superar este estancamiento requiere compromiso de la alta dirección ya que como dice Joseph Juran "El liderazgo de alta calidad no se puede delegar".

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).

Valeria Reyes Tapia

El objetivo del artículo es hacer visible que desarrollar una cultura para la recolección de datos precisos y de calidad los empleados necesitan ver a los líderes organizacionales involucrados activamente en el proceso y aceptar la responsabilidad total por la calidad de los datos dentro de sus respectivas unidades de negocios. Una vez que esto se haya logrado, la calidad de los datos será responsabilidad de todos y no solo de TI.

Ya que hoy en día el mayor reto para las organizaciones es limpiar sus procesos de administración de datos ya que recordemos que la calidad de datos está ligada a la recolección total de información dentro de un sistema o almacén de datos y al poseer datos sin calidad puede representar una desventaja y al querer trabajar rápidamente para recopilar datos y usarlos para optimizar

programas casi en tiempo real, nos lleva fácilmente a depender de datos inexactos, incompletos o redundantes, creando un efecto dominó de decisiones basadas en números y métricas inexactas.

Todo esto es de gran importancia ya que permite garantizar el procesamiento masivo de datos para contribuir a que los resultados que arrojen sean óptimos y contribuyan a la correcta toma de decisiones.

Estoy totalmente de acuerdo de comenzar con cultura a tiempo de gestión de datos, ya que con esto obtendríamos datos precisos y actualizados, los cuales son esenciales y facilitan la toma de decisiones sólidas en las empresas. El más claro ejemplo desde mi punto de vista se da en el trabajo de equipo, ya que cuando diferentes departamentos tienen acceso a datos coherentes, es más fácil para las empresas mantenerse alineadas en cuanto a prioridades, o para generar resultados más estratégicos y cohesivos.

Santiago Díaz:

En este artículo, se trata de evidenciar un problema que es muy común cuando se manejan recolecciones de datos; la **recolección precisa de datos** , ya que un error en la recolección hace que estos datos sin **calidad**, e inclusive que haya consecuencias peligrosas.
Esto se puede arreglar sin necesidad de invertir en nuevas tecnologías ni nada por el estilo, basta con que los creadores de los sistemas de recolección se comuniquen con aquellos que recolectan datos. Usualmente, cuando suceden problemas en la calidad de la recolección, se le asigna la responsabilidad a IT de solucionarlo, pero usualmente esto no funciona, ya que la calidad de los datos es solucionada a la hora de que se crean, es por esto que la comunicación entre recolectores y personal de IT puede solucionar este problema.
De este artículo, debemos entender que la **comunicación** entre recolectores y personal de IT debe de ser prioritario para tener una calidad en los datos en una empresa. Se da el ejemplo de una empresa millonaria, *Chevron*, donde se dieron cuenta que había impresición en sus datos. Dada esta impresición, contrataron a expertos en manejo de datos para solucionar su problema, y un punto muy importante a recalcar es que antes que esforzarse en corregir los datos imprecisos, las compañías deberían mantenerse concentradas en mejorar la forma en que crean nuevos datos, de esta manera lograron "limpiar" los datos y tener calidad en sus recolecciones en menos de un año.
Desde mi punto de vista, el autor tiene toda la razón en su forma de solucionar estos problemas. Debemos de tener una mayor comunicación entre áreas de una empresa para evitar impresiciones.

David Hernández Uriostegui

El propsito inicial de este árticilo tiene como objetivo darnos a conocer y entender las repercuciones que pueden llegar a suceder por el uso y/o creación de datos de baja calidad, y que este tipo de problemas no son solamente de un area en especifico, si no que en general el uso de este tipo de datos impacta de diferentes maneras, pero negativamente.

Adicionalemente como se nos mencionó en clase, las personas que trabajan en obtener gastan la mitad de su tiempo laboral limpiando y buscando datos para evitar hacer analisis o labores de baja calidad.

Inmediatamente se nos proporciona un ejemplo de como es que los datos dentro de una empresa se vuelven de baja calidad, siendo afectados no solamente el equipo que recavó los datosm si no que todos los equipos que hacen uso de estos datos se ven afectados provocando que surgan nuevos inconvenientes como el gasto de capital más elevado de lo que debería haber sido, este tipo de cosas se nos han mencionado desde las primeras clases que para poder evitar/solucionar la creación de

datos de baja calidad es fundamental la comunicación entre todos los involucrados.

El resto de la lectura se abordan el problema de como resolver el problema de la creación y limpieza de datos de baja calidad, y el primer para esto es haya comunicación entre los usuarios de datos y los creadores de datos, para que de esta manera se puedan fijar nuevos objetivos sobre qué tipo de datos crear y como organizarlos.

El siguiente paso es *"limpiar"* los datos de mala calidad que ya se tiene guardados, pero en realidad el limpiar estos datos no va a generar un cambio inovador y beneficioso para la organización, lo que se debe hacer en este caso para no gastar tiempo y enormes cantidades de dinero, es diseñar la estructura de los nuevos datos para que de esta forma la nueva base de datos esté siendo actualizada continuamente y que las limpiezas que deban hacerse sean limitadas, este enfoque es mucho mejor tanto como en cuestión ecónomica como técnologicamente.

Y este tipo de enfoque se rectifica mostrandonos un ejemplo dónde una empresa decidió seguir esta orientación, y al hacerlo la empresa recibió bastantes beneficios, y de hecho el autor mismo menciona que el ha visto en diversas ocasiones que al seguir esta perspectiva para atacar el problema de calidad de datos siempre muestra resuldados similares (mejoramiento de la calidad de datos).

Personalmente creo que los temas abordados por el árticulo son las principales problematicas al lidiar con bases de datos, desde su creación y planificación hasta la limpieza de estas, cada uno de estos aspectos tiene sus problematicas y repercusiones.

Yo concuerdo con todo lo mencionado en el árticulo, considero que para poder generar datos que seen de confianza y calidad debe haber comunicación constante con los usuarios del producto en el cuál se está trabajando para que se pueda tener bases de datos estructuradas de acuerdo a las princpales necesidades y que permitan que la limpieza de estas sea limitada, y en caso que se deba hacer una limpieza profunda de la base datos, lo mejor será volver a restructurar la manera en que se crean los datos para evitar costos elevados y poder trabajar de manera más rapida y eficiente con datos de calidad buena.

Diego J. Padilla Lara

Es muy claro que la recolección y manejo de datos no es un proceso del todo perfeccionado. Todos los días se genera una gran cantidad de "basura" (datos imprecisios, incorrectos o simplemante inservibles) que no sólo entorpece el funcionamiento de la industria, si no que además representa pérdidas económicas significativas. Si bien las empresas se las han arreglado para trabajar de ésta manera, también es cierto que han pasado cincuenta años desde que la frase *"garbage in, garbage out"*, que hace alución a la mala calidad de los datos (y lo que traen consigo), se volvió un recurrente en la industria.

Se vuelve alarmante el saber que a pesar de no ser un problema reciente, no se ha conseguido mejorar la calidad de los datos. Lo cuál nos hace preguntar ¿existirá alguna forma de mejorar éste proceso?

Para responder a la pregunta, primero necesitamos ver los datos desde un punto de vista enfocado en calidad y centrarnos en su ciclo de vida. Vemos que hay dos momentos relevantes: cuándo se generan, y cuándo se utilizan. Suponiendo que los datos no tienen alteración al momento de ser utilizados, se hace obvio que la calidad de éstos depende completamente del momento en el que se

generan.

Claramente para mejorar todo el proceso hace falta atender el momento en el que se recolectan los datos. Esto no es algo que se pueda conseguir de la noche a la mañana pues requiere que las empresas modifiquen sus modelos de trabajo al distribuir la responsabilidad de generar datos de calidad sobre más departamentos que sólo el de IT.

Los pasos a seguir puden ser resumidos en:

- Conectar a los creadores con los consumidores de datos. Así se identifican problemas y se consigue que los creadores pongan más atención al momento de generar los datos, pues saben con que fines se estarán utilizando.
- Concentrarse en que los nuevos datos obtenidos sean de calidad. Más que tratar de corregir todos los datos antigüos.

En mi opinion, esta problemática no puede ser resuelta del todo porque siempre existirá el error humano; sin embargo, éste margen de error puede ser disminuido drásticamente al prestar más atención al proceso de obtención de datos.