Almacenes y minería de datos Flores González Luis Brandon
Semestre 2018-II
Tarea 1: Conceptos generales y diseño

1. **Describe al menos 6 características de los DWH.**

    a. Orientada a un tema. Tema como cliente, proveedor, producto, venta. En lugar de procesamiento de transacciones de una organización.
    b. Integrada. Usualmente se construye integrando múltiples fuentes heterogéneas. Se requieren técnicas de limpieza e integración de datos para asegurar la consistencia entre los datos.
    c. Históricos. Los datos se almacenan para proporcionar información desde una perspectiva histórica. Cada elementos clave contiene explícita o implícitamente un elemento de tiempo.
    d. No volátil. No requiere mecanismos para procesamiento de transacciones, recuperación y control de concurrencia. Sólo requiere dos operaciones para accesar los datos: carga iniciar y acceso de datos.
    e. Usa procesamiento analítico en línea(OLAP). Menos consultas pero más grandes, lecturas frecuentes, actualizaciones frecuentes(diariamente, semanalmente), operaciones de lectura o actualización(en dos faces), grandes volúmenes de datos(colección de datos históricos), modelo de datos sencillo(multidimensional/de-normalizado)
    f. Calidad de datos. Diferentes fuentes típicamente utilizan diferentes representaciones, códigos y formatos de datos que deben ser unificados. Y así garantizar una mejor calidad en los datos.

2. **¿Porqué consideras que es necesaria la integración de los datos en un DWH y no en una aplicación de BD?**

    Ya que esta toma los datos de diferentes fuentes heterogéneas, lo que nos permite hacer un proceso de limpiar los datos y tener un calidad de ellos, además de transformarlos. Para finalmente obtener información de estos, que es la parte que interesa para la toma decisiones. Si no integramos estos datos en un DWH tendríamos mala calidad de datos, las fuentes serían heterogéneas lo cual tendría malas repercusiones para obtener la información ya que hay dispersión por las diferentes fuentes, los datos serían volátiles.

    Esto nos lleva a tener un alto rendimiento para los propósitos que tenemos, los cuales son: consultas complejas, vistas multidimensionales y consolidación.

3. **Describe los tres principales tipos de metadatos que se encuentran en un DWH.**
    a. **Operacionales:** se refieren a los metadatos generados y capturados cuando se ejecuta un proceso. Permite que los administradores gestionen su sistema y aseguran que las cosas funcionen sin problemas. Si hay un problema con algún proceso, los metadatos operacionales también ayudan a los administradores a identificar y localizar los problemas.
    Ejemplos: información acerca de la ejecución de las aplicaciones, incluyendo la frecuencia, conteos de registro, un análisis de componente por componente y otras estadísticas con fines de auditoría.
    b. **Extracción y transformación:** describen la despensa o el almacén de datos de destino.
    c. **Para el usuario final:** ayudan al usuario a acceder al almacén de datos con su propio lenguaje de negocio, indicando qué información hay y qué significado tiene. Ayudar a construir consultas, informes y análisis, mediante herramientas de Business Intelligence como DSS, EIS o CMI.
4. **Explica en tus propias palabras, qué es la arquitectura de un DWH y cómo funciona.**

    Son los diferentes entornos por lo que pasan los datos desde las fuentes hasta un almacén de datos, estos son:

a. **Fuentes de datos:** es el origen de los datos, estos pueden ser sistemas OLTP(sistemas que son diseñados para trabajar de forma independiente), archivos de texto, sistemas heredados, hojas de cálculo, archivo en papel, etc.

b. **Preparación de datos:** es el área intermedia donde se realizan la transformación, integración y limpieza de los datos. Estos procesos se llevan a cabo entre las fuentes y el almacén de datos. Esto es con el fin de mejorar y asegurar la calidad de los datos.
   Además aquí se encuentra el monitor, se almacena los datos y tiene la capacidad de re-cargar los datos que llegaron a esta etapa. Dentro de esta etapa se encuentra los siguientes pasos:
   i. **Monitor:** determina cambios en los datos. Captura los cambios en el contenido de los datos de los sistemas de origen(no suele hacer hacerse en la carga inicial pero si posteriormente). En otras palabras, su objetivo es descubrir cambios en las fuentes de datos de forma incremental.
   ii. **Extracción:** Se identifica las fuentes de datos a las cuales se les realizara los procesos ETL para después ser procesados.
   iii. **Transformación:** convierte los datos en algo que sea representable y con valor para el negocio, esto involucra el análisis de: Limpieza de Datos, Datos no existentes, Datos extremos e Integración de esquemas.
   iv. **Carga:** almacena los datos de forma rápida en el DWH. Ya sea actualización ó carga masiva.

5. **Describe los conceptos: esquema estrella, copo de nieve y constelación.**

   a. **Esquema estrella:** es una estructura que consta de una tabla central de hechos y varias dimensiones, estas están relacionadas a la tabla de hechos. Lo característico de esta arquitectura es que sólo existe una tabla de dimensiones para cada dimensión.
      En otras palabras, la única tabla que tiene relación con otra es la de hechos, lo que significa que toda la información relacionada con una dimensión debe estar en una sola tabla.
   b. **Copo de nieve:** es una variación o derivación del modelo estrella, en esta estructura la tabla de hechos deja de ser la única relaciona con otras tablas ya que existen otras tablas que se relacionan con las dimensiones y que no tienen relación directa con la tabla de hechos. Este modelo hace que la extracción de datos sea más difícil así como vuelve compleja la tarea de mantener el modelo.
   c. **Constelación:** es una combinación de un esquema de estrella y un esquema de copo de nieve. Esta estructura son esquemas de copo de nieve en los que sólo algunas de las tablas de dimensiones se han desnormalizado.
      Las jerarquías de los esquemas de estrella están desnormalizadas, mientras que las jerarquías de los esquemas de copo de nieve están normalizadas.
      Para normalizar el esquema, las jerarquías dimensionales compartidas se colocan en outriggers(entidad unida a otras tablas de dimensiones).

6. **Un DWH es orientado a un tema. ¿Cuales podrán ser los aspectos críticos en las siguientes organizaciones?**
   a. **una compañía manufacturera internacional.** Las ventas(hecho), producto, locación de manufacturación, tiempo, comprador, vendedor.Estas últimas como dimensiones.
   b. **un banco de una comunidad local.** Cuentahabiente(hecho), saldo, tiempo, transacción, etc. Estas últimas como dimensiones.
   c. **una cadena hotelera nacional.** Ventas(hecho), tiempo, hotel, perfil del huésped, etc. Estos últimos como dimensiones.

   Cabe aclarar, que el hecho es el enfoque del análisis mientras las dimensiones son los factores por lo que se analiza un determinado área del negocio. Y esto viene dado de lo que se quiere analizar.

7. **Lee el articulo "Data Cleaning: Problems and Current Approaches" que se encuentra en la página del curso en la sección de Material/Lecturas. Responde las siguientes preguntas:**

   a. **¿Qué es la limpieza de datos?**
   Se detectan y remueven errores e inconsistencias desde los datos para proveer la calidad de datos.

   b. **¿Cuál es el objetivo de la limpieza?**

   El objetivo es detectar y eliminar todos los errores e incoherencias importantes tanto en las fuentes de datos individuales como al integrar múltiples fuentes. El enfoque debe ser respaldado por herramientas que limiten la inspección manual y el esfuerzo de programación y sean extensibles para cubrir fácilmente fuentes adicionales. Además, la limpieza de datos no se debe realizar de forma aislada, sino junto con transformaciones de datos relacionadas con esquemas basadas en metadatos completos.

   c. **¿Qué significa Calidad de Datos?**

   Es una evaluación de la utilidad de los datos para cumplir su propósito en un contexto determinado.

   d. **¿Qué significa Gobierno de Datos?**

   La calidad de los datos de una fuente depende en gran medida del grado en que se rige por las restricciones de esquema e integridad que controlan los valores de datos permisibles.

   e. **¿Cuáles son los problemas que enfrenta hoy en día la Limpieza de Datos?**

   Un problema principal para limpiar datos de múltiples fuentes es identificar datos superpuestos, en particular registros coincidentes que se refieren a la misma entidad del mundo real. Consecuentemente se ven reflejados como problemas de calidad de datos, estos se pueden clasificar como sigue:
      1. Problemas de fuente única:
         a. Nivel de esquema: falta de integridad, restricciones, pobres diseño del esquema.
         b. Nivel de instancia: errores de entrada de datos.
      2. Problemas de varias fuentes:
         a. Nivel de esquema: modelos de datos heterogéneos y diseños de esquema.
         b. NIvel de instancia: superposición, contradicción y datos inconsistentes.

   f. **¿Qué enfoques aborda para solventar dichos problemas?**

   Como la limpieza de las fuentes de datos es un proceso costoso, la prevención de la entrada de datos sucios es obviamente un paso importante para reducir el problema de limpieza.

   g. **¿Qué es el análisis de datos y cómo se puede utilizar para apoyar las tareas de limpieza de datos?**

   Es una evaluación de la utilidad de los datos para cumplir su propósito en un contexto determinado.

   h. **¿De qué forma los procesos ETL ayudan a efectuar la Limpieza de Datos?**

En primer lugar son múltiples pasos donde cada uno quizá realice transformaciones(mapeo) de esquema y relaciones de instancia. Para permitir una transformación de datos y limpiar el sistema y por lo tanto reducir el monto de autoprogramación es necesario especificar las transformaciones requeridas en un **lenguaje apropiado**(como algunas herramientas ETL). Una forma más general y flexible es el uso del **lenguaje SQL** para realizar las transformaciones de datos y utilizar la posibilidad de especificar y aplicar las extensiones de lenguaje, en particular las **funciones definidas**, con estas funciones se pueden aplicar una gran cantidad de transformaciones para diferentes tareas de transformación y procesamiento de consultas.Estas funciones definidas aun implican un esfuerzo de implementación y no soportan todas las transformaciones de esquemas necesarios. Frecuentemente se necesitan funciones tal como división de atributos o mezclados que generalmente no son soportados pero se necesita aveces **re-implementar** en las aplicaciones.
Para transformaciones de esquemas relacionados, las extensiones de lenguaje tal como **SchemaSQL** son requeridos.

i. ¿Qué mecanismos propondrías para eliminar o minimizar el impacto de la mala calidad de los datos?

Principalmente, creo que la forma más fácil de minimizar la mala calidad de datos es controlar el flujo de datos en el origen de las fuentes, es decir, la forma en que se almacenar o controlan los datos en estas fuentes. Teniendo un mayor control o restricciones desde el origen, para evitar varias transformaciones y tener una calidad de datos desde el comienzo.

j. Conclusiones generales sobre el artículo.

Finalmente la limpieza de datos es un punto muy importante que se debe considerar en todo momento ya que de esta depende la información que obtendremos en el análisis. Para lograr esta calidad de datos impredecible, se puede lograr con varias herramientas o transformaciones que en ningún momento se debe olvidar ya que se pueden obtener varios problemas con esta calidad. En pocas palabras, la limpieza de datos es un gran esfuerzo que se requiere hacer para lograr eficiencia, veracidad, precisión, etc.