GARDUÑO VARGAS JUAN CARLOS. TAREA 1

1. Describe al menos 6 características de los DWH.
   - Orientado a temas
   - Integrado
   - No volatil
   - De tiempo variante
   - Accesibilidad
   - Seguridad

2. ¿ Por que consideras que es necesaria la integración de los datos en un DWH y no en una aplicación de BD?
   - Para evitar la redundancia en los datos, ya que la integración de los datos facilita que diferentes tipos de datos sean usados por usuarios, organizaciones, etc.

3. Describe los tres principales tipos de metadatos que se encuentran en un DWH.
   - Metadatos operacionales: Se refieren a los metadatos generados y capturados cuando se ejecuta un proceso.
   - Metadatos extraccion y transformacion: Identifica los datos que han cambiado y hace la estandarización a un solo formato.
   - Metadatos Usuario final: Ayuda a acceder al Data Warehouse con su propio lenguaje de negocio, indicando qué información hay y qué significado tiene. Ayudar a construir consultas, informes y análisis, mediante herramientas de navegación.

4. Explica en tus propias palabras, que es la arquitectura de un DWH y como funciona.
   - Es un conjunto de datos que a diferencia de una base de datos, esta está formada por variables y dimensiones. Donde su objetivo es ayudar a la toma de decisiones con la capacidad de realizar en tiempo real análisis multidimensionales.

5. Describe los conceptos:

   - Esquema estrella: Es la arquitectura de almacén de datos mas simple. En este diseño el almacén de datos la tabla de Variables esta rodeada por dimensiones y juntos forman una estructura que permite implementar mecanismos básicos para poder utilizarla con una herramienta de consultas OLAP.

   - Esquema Copo de Nieve: Es una variedad mas compleja del esquema estrella, ya que el afinamiento esta orientado a facilitar mantenimiento de dimensiones ademas de que las tablas de dimensiones en este modelo representan relaciones normalizadas y forman parte de un modelo relacional de base de datos.

   - Esquema de Constelación: Para cada esquema mencionado anteriormente se puede construir un esquema de constelación de hechos. Este esquema es mas complejo debido a que contiene múltiples tablas de hechos.

6. Un DWH es orientado a un tema. ¿Cuales podría ser los aspectos críticos en las siguientes organizaciones?

- Una compañía manufacturera internacional: Uno de los problemas que podria tener es que pueda tener como objetivo las ganancionas, el tipo de producto que maje la empresa, las cantidades, etc.

- Un banco de una comunidad local: Asi mismo se puede tener un contraste entre dos tipos de orientaciones como lo son los orientados a las aplicaciones y los orientados a temas. Para este caso un banco local deberia procurar tener como objetivo el cliente, vendedor, la actividad, etc. En vez de los ahorros, depositos, prestamos.

- Una cadena hotelera nacional: Como lo hemos manejado, para una cadena hotelera, esta deberia enfocarse en los clientes que se hospedan, cuartos disponibles, actividad frecuente, trabajadores, etc.

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

(a) ¿Que es la limpieza de datos?

-La limpieza de datos llamada data cleansing o scrubbing es la que se ocupa de detectar y eliminar errores e inconsistencias de los datos con el fin de mejorar la calidad de los datos.

(b) ¿Cual es el objetivo de la limpieza?

- La limpieza de datos debe cumplir varios requisitos. En primer lugar, debe 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) ¿Que significa Calidad de Datos?

- La calidad de los datos es una evaluación de la utilidad de los datos para cumplir su propósito en un contexto determinado.

(d) ¿Que 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) ¿Cuales son los problemas que enfrenta hoy en dıa 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. Este problema también se denomina problema de identidad del objeto, eliminación duplicada o problema de combinación. Con frecuencia, la información es solo parcialmente redundante y las fuentes pueden complementarse proporcionando información adicional sobre una entidad.

(f) ¿Que enfoques aborda para solventar dichos problemas?

- Dado que 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. Esto requiere un diseño apropiado del esquema de la base de datos y las restricciones de integridad, así como de las aplicaciones de entrada de datos.

(g) ¿Que es el análisis de datos y como se puede utilizar para apoyar las tareas de limpieza de datos?

- Para poder detectar que tipos de errores e inconsistencias deben eliminarse, se requiere un análisis de datos detallado. Además de una inspección manual de los datos o muestras de datos, los programas de análisis deben utilizarse para obtener metadatos sobre las propiedades de los datos y detectar problemas de calidad de datos.

(h) ¿De que forma los procesos ETL ayudan a efectuar la Limpieza de Datos?

- Una forma 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 y reutilización para diferentes tareas de transformación y procesamiento de consultas.

(i) ¿Que mecanismos propondrías para eliminar o minimizar el impacto de la mala calidad de los datos?

- Una forma seria validar la exactitud e integridad de los datos desde la fuente para evitar que los datos incorrectos se sigan moviendo por los sistemas.

- Otra forma seria buscar los problemas mas importantes y tratar de resolverlos por partes.

(j) Conclusiones generales sobre el articulo.

- El articulo maneja diversos problemas así como es la calidad de los datos y los problemas del nivel de instancia y de esquema. Ademas se explica los pasos para lo que es la transformación y la limpieza de datos, brindando descripciones de las herramientas que generalmente cubren solo una parte del problema. En lo personal es un articulo que explica muy bien los temas que son el análisis de datos, los enfoques que tiene la limpieza y como hemos mencionado, los problemas que conlleva por lo que todos estos temas son de gran ayuda para complementar el curso.