TRABAJO PRÁCTICO FINAL

Comentarios iniciales

- El siguiente trabajo práctico debe ser realizado individualmente o en grupos de hasta 2 personas .
- Requiere instalación de software en equipo propio o bien uso de laboratorio
- Es posible consultar sobre este trabajo a los docentes de las cátedras por los canales ya establecidos. Recomendamos lectura atenta al enunciado.
- Para subir al campus virtual la entrega de este TP favor de generar un archivo .zip con este formato de nombre `BBDD2-COM1-TPFinal-ApellidosNombres.zip`.
- Tiempo estimado para resolución: 8 bloques de 4hs.

Recursos

Para la realización de este Trabajo Práctico deberá emplear el Software especificado en el documento Software requerido

Desarrollo del Trabajo Práctico

Este documento consta de:

- Situación de Negocio
- Descripción de las fuentes de datos a utilizar
- Consultas requeridas por el departamento de Ventas

Situación de Negocio

Ud. es un administrador de base de datos para una compañía llamada The Drinking Company (TDC). La compañía produce bebidas de distintas categorías que comercializa tanto en el mercado minorista (retail) como en el mercado mayorista (wholesale) y trabaja los 365 días del año.

TDC tiene una trayectoria de 4 años en el mercado con un gran éxito en el rubro. Posee un área de ventas que abarca varios estados de los Estados Unidos, repartidos en 4 regiones (east), oeste (west), central (central) y sur (south).

Los productos que comercializa TDC están divididos en 5 rubros: colas (cola), cervezas (beer), sodas (soda), jugos (juice) y bebidas energizantes (energy drink). Dichos productos pueden tener varias presentaciones: botellas de 1 litro, botellas de 2 litros, botellas de 670 cm^3, latas de 330 cm^3 y latas de 500 cm^3.

El departamento de finanzas quiere rastrear, analizar y pronosticar el ingreso de las ventas, a través de las zonas geográficas, los productos, los clientes y diferentes períodos de tiempo.

Ud. ya ha definido algunas consultas básicas. Sin embargo, estas consultas agregan trabajo a la base de datos operacional. Además, los usuarios solicitan consultas adicionales a medida que las necesitan.

Su compañía ha decidido crear un Data Mart para la información de ventas. Un Data Warehouse contiene información que ha sido limpiada y transformada a un formato “informativo” y no operacional.

Descripción de las fuentes de datos a utilizar

La Empresa cuenta con datos distribuidos en distintos sistemas operacionales:

Area de Recursos Humanos

Datos de los empleados y sus vacaciones. Este sector mantiene la información en archivos de Excel.

Holidays.xls (Vacaciones)

CAMPO
DATE
HOLIDAY

Employee.xls (Empleados)

CAMPO
EMPLOYEE_ID
FULL_NAME
GENDER
CATEGORY
EMPLOYMENT_DATE
BIRTH_DATE
EDUCATION_LEVEL

Area de comercialización

El área de comercialización provee información de los clientes de la compañía a los distintos subsistemas mediante tecnología de web services.
Esta información se nos ha suministrado en dos archivos XML: uno para clientes minoristas (customer_R.xml) y otro para clientes mayoristas (customer_W.xml) Los campos contenidos en los archivos XML son los siguientes:

Customer_R.xml (Clientes Minoristas)

CAMPO
CUSTOMER_ID
FULL_NAME
BIRTH_DATE
CITY
STATE
ZIPCODE

Customer_W.xml (Clientes Mayoristas)

CAMPO
CUSTOMER_ID
FULL_NAME
BIRTH_DATE
CITY
STATE
ZIPCODE

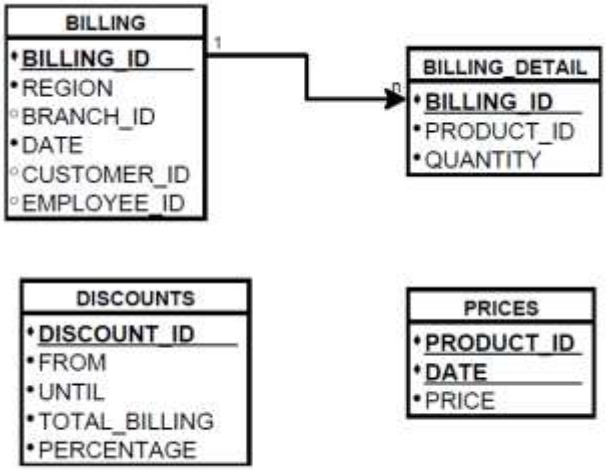
Este sector nos proveerá también la información de las ventas actuales, y lasventas históricas.

Los datos de las ventas estuvieron almacenados en una Base de Datos SQL Server 2000 hasta el año 2008 inclusive.
A partir de enero de 2009 se decidió realizar una migración a MySQL. Esta base de datos incluye una tabla con los precios de los productos a lo largo del tiempoy otra con los descuentos aplicados a las ventas. El precio de un producto se identifica con el código del producto y la fecha y hora en la que comenzó a tener vigencia dicho precio. A una factura se le aplica un descuento si el monto total dela factura supera o iguala el monto asociado a un descuento y si además ese descuento rige para esa fecha. En el caso de que sean factibles más de un descuento se aplica el mejor de ellos (el de mayor porcentaje).

Base de datos SQLSERVER 2000 de ventas históricas (history sales)

BILLING
ID
BILLING_ID
DATE
CUSTOMER_ID
EMPLOYEE_ID
PRODUCT_ID
QUANTITY

Base de Datos MySQL de ventas actuales (Sales)

- 

-
-
-

Se nos provee también el archivo de regiones, en formato de texto plano.

Regions.txt (Regiones)

CAMPO
REGION
STATE
CITY
ZIPCODE

Area de producción

El sector de producción, opera actualmente con un potente sistema MRP, y nosha suministrado la información de los productos y los movimientos de stock enformato de texto plano.

Products.txt (Productos)

CAMPO
PRODUCT_ID
DETAIL
PACKAGE

Stock.txt

CAMPO
PRODUCT_ID
DATE
VARIATION

Consultas requeridas por el departamento de Finanzas

Se solicita la construcción de un Data Mart que atienda a las siguientes cuestiones:

1. Cantidad de litros consumidos y de productos adquiridos por cliente en eltiempo

2. Compra promedio en litros por cliente en el tiempo

3. Rankear los productos por zonas geográficas a través del tiempo

4. Suponiendo que las condiciones económicas se mantengan aproximadamente estables, es posible predecir el monto de las ventas para el primer trimestre del año 2011?

5. El gerente de Marketing desea preparar una promoción de importantes descuentos en las bebidas tipo Energy Drink para promocionar este tipo de bebidas en los eventos deportivos a producirse en los meses de setiembre, porque piensa que coincide con una etapa de picos en el monto de ventas dentro del año. Es correcta esta afirmación?

6. El gerente de Marketing también quiere saber cómo es la relación entre lasedades y los tipos de bebida, teniendo en cuenta la cantidad de litros vendidos. Es importante el tipo de bebida en la determinación de los grupos etarios?

7. También para la segmentación de los consumidores se desea saber puntualmente cuál es el monto de ventas global para los teenagers (13-19), para los adultos medios (40-50) y, por un capricho propio del gerente, para los consumidores de su misma edad (66 años)

8. El gerente de RRHH necesita saber si la edad y el sexo del empleado tienen relación con el monto de ventas.

9. También desea saber cuáles serán los 5 vendedores más prometedores enmonto de ventas para el año 2011.

10. Se desea saber también si los vendedores con mayor antigüedad en elempleo venden la mayor cantidad de productos.

11. Se desea saber cuáles son los clientes minoristas más valiosos siguiendo elprincipio de Pareto.

12. El gerente de RRHH necesita saber si la edad y el sexo del empleado tienen relación con el monto de ventas.

13. El gerente supone que las bebidas diet están perdiendo popularidad.

14. Se desea saber tambien si las bebidas en lata están bajando su consumo

15. Determinar que productos discontinuar.

16. Determinar cuáles son los clientes más valiosos para la empresa y sucomportamiento a lo largo del tiempo

Los puntos a desarrollar en el Trabajo Practico son los siguientes:

- **Diagrama estrella del Data Warehouse:** Utilizando los conceptos aprendidos, crea un diagrama estrella que represente la estructura lógica de tu Data Warehouse (detallar tabla de hechos , dimensiones , sus relaciones y los atributos de las mimas, marcar PK y FK, cardinalidad y jerarquías si las hubiera)

- **Definir el flujo de datos y las tareas para creación del Data mart**

- **Implementación del Data Warehouse**: Basado en el punto anterior, crear una base de datos para el Data Warehouse utilizando SQL Server.

- **Carga del Data Warehouse**: Describe detalladamente los pasos necesarios para cargar correctamente el Data Warehouse. Incluye la extracción de datos desde diferentes fuentes, su transformación y limpieza, y finalmente, la carga en las tablas del Data Warehouse. Se deberá contemplar:

    - Carga de datos desde las bases de datos operativa de la empresa

    - Carga de datos desde fuentes externas de ser necesario

- **Implementación de ETL en SSIS**: Utilizando SQL Server Integration Services (SSIS), implementa los pasos descritos anteriormente para cargar el Data Warehouse. Asegúrate de mostrar un flujo de trabajo lógico y eficiente utilizando las herramientas proporcionadas por SSIS.

- **Creación de Dashboards para la gerencia**: Utilizando PowerBI u otra herramienta de visualización de datos de tu elección desarrollar los tableros y reportes necesarios

Estandares de nomenclatura al realiza el TP de BD2

Nombre de Bases de datos :

bd_staging_YYYY_Gxx
bd_intermedia_YYYY_Gxx
datawarehouse_YYYY_Gxx
Donde YYYY es el Año de Cursada de la materia y xx es el numero de grupo asignado

Nombre de tablas :

-Las tablas de la base bd_staging_YYYY_Gxx deberan comenzar con stg_xxxxx_Gxx (Ejemplo stg_clientes_G01)

-Las tablas de la base bd_intermedia_YYYY_Gxx deberan comenzar con int_xxxxx_Gxx (Ejemplo int_clientes_G01)

-Las tablas de la base datawarehouse_YYYY_Gxx si son dimensiones Dim_xxxxxxx_Gxx (Ejemplo Dim_clientes_G01) si son tablas de hechos Fact_xxxxxxx_Gxx (Ejemplo Fact_ventas_G01)

-El nombre de las tablas debe ser en mayúsculas o CamelCase.

-El nombre de las tablas deben ser descriptivos, no importa que tan largos sean siempre y cuando sean soportados por la base de datos.

-Si la tabla tiene más de 2 palabras estas se deben poner juntas o con un guión bajo, nunca se debe de usar espacios, ej 'APELLIDO_PATERNO' o en CamelCase 'ApellidoPaterno'.

Nombre campos :

Campos en base de datos datawarehouse_YYYY_Gxx (en base datastaging e intermedia no es necesario)

El nombre de los campos deben de ser CamelCase, empiezan en minúsculas, no tienen espacios o guiones bajos, son descriptivos, y las siguientes palabras empiezan con mayúscula, ej 'holaMundo', 'apellidoPaterno', sin embargo también pueden ir en mayúscula usando un guión bajo como espacio, la única condición es que todos sean homogéneos y claros .

Convenciones de nomenclatura ETLs :

Utilizar nombres descriptivos en los orígenes y destinos de datos, tablas, columnas, variables, funciones y procedimientos. Las convenciones de nomenclatura deben ayudar a evitar confusiones y ambigüedades, y hacer que el código sea más legible y comprensible. Debe usar nombres descriptivos y significativos que reflejen la lógica de negocios y el flujo de datos de los patrones de diseño de ETL. También debe utilizar prefijos o sufijos para indicar el tipo, la función o el estado de los elementos de datos, como el origen, el destino, el almacenamiento provisional, la búsqueda, el error o el archivo.

Puntos Extra:

- Actualización Incremental por Fecha : Considerando la necesidad de actualizar el Data Warehouse de forma incremental según la fecha de facturación, analiza y describe las modificaciones que tendrías que realizar en los pasos anteriores. Explora cómo identificar las novedades , actualizar las tablas correspondientes y garantizar la coherencia de los datos existentes y nuevos.

- Reportes Adicionales: Cualquier reporte adicional a los solicitados en el TP#1 será contemplado como puntaje adicional.

Este trabajo práctico representa una excelente oportunidad para profundizar en el diseño, carga y visualización de datos en un entorno empresarial. ¡Explora nuevas técnicas y herramientas, y desarrolla habilidades relevantes para el mundo del Data Warehouse !