Universidad Simón Bolívar
Departamento de Computación y T.I.
CI-5313 Arquitectura y administración de bases de datos
Abril-Julio 2016

Construcción de la Base de Datos de FBV

La Fábrica de Bombas de Venezuela FBV ofrece sus servicios a través de Internet a distribuidores y particulares, para ventas en grandes volúmenes, por productos o partes de repuesto individuales. FBV es lo que se considera actualmente como un "on-line supplier".

Actualmente FBV está estudiando el desarrollo de un "Data Warehouse" para evaluar y analizar sus ventas, para lo cual se van a realizar consultas sobre una copia de la base de datos operativa con la finalidad de realizar actividades de minería de datos ("data mining") sobre la copia. Esta copia de la base de datos que se va a utilizar para análisis se va a refrescar (refresh) periódicamente con nuevos datos de la base de datos operativa o eliminando algunas tuplas de acuerdo con algún criterio.

FBV registra información de sus clientes en la relación CUSTOMER, la cual incluye el país; los países se encuentran codificados en la relación NATION y la región en la cual se encuentra ese país se representa en la relación REGION. Los productos que vende FBV son bombas de agua y sus partes, los cuales se encuentran descritos en la relación PART. Además, existe una relación SUPPLIER que contiene los fabricantes de los productos, para los cuales se describe también el país y la región donde se encuentran sus oficinas principales. Cada fabricante puede proveer muchos productos y viceversa, lo cual se refleja en la relación PARTSUPP. Las órdenes de cada cliente se almacenan en la relación ORDERS y los detalles de las órdenes se almacenan en la relación LINEITEM. Cada línea de detalle de una orden contiene un producto de un fabricante específico. En la sección 1 se presenta el modelo lógico de la base de datos de la empresa con toda la información detallada de la implementación de las relaciones en el manejador de base de datos seleccionado.

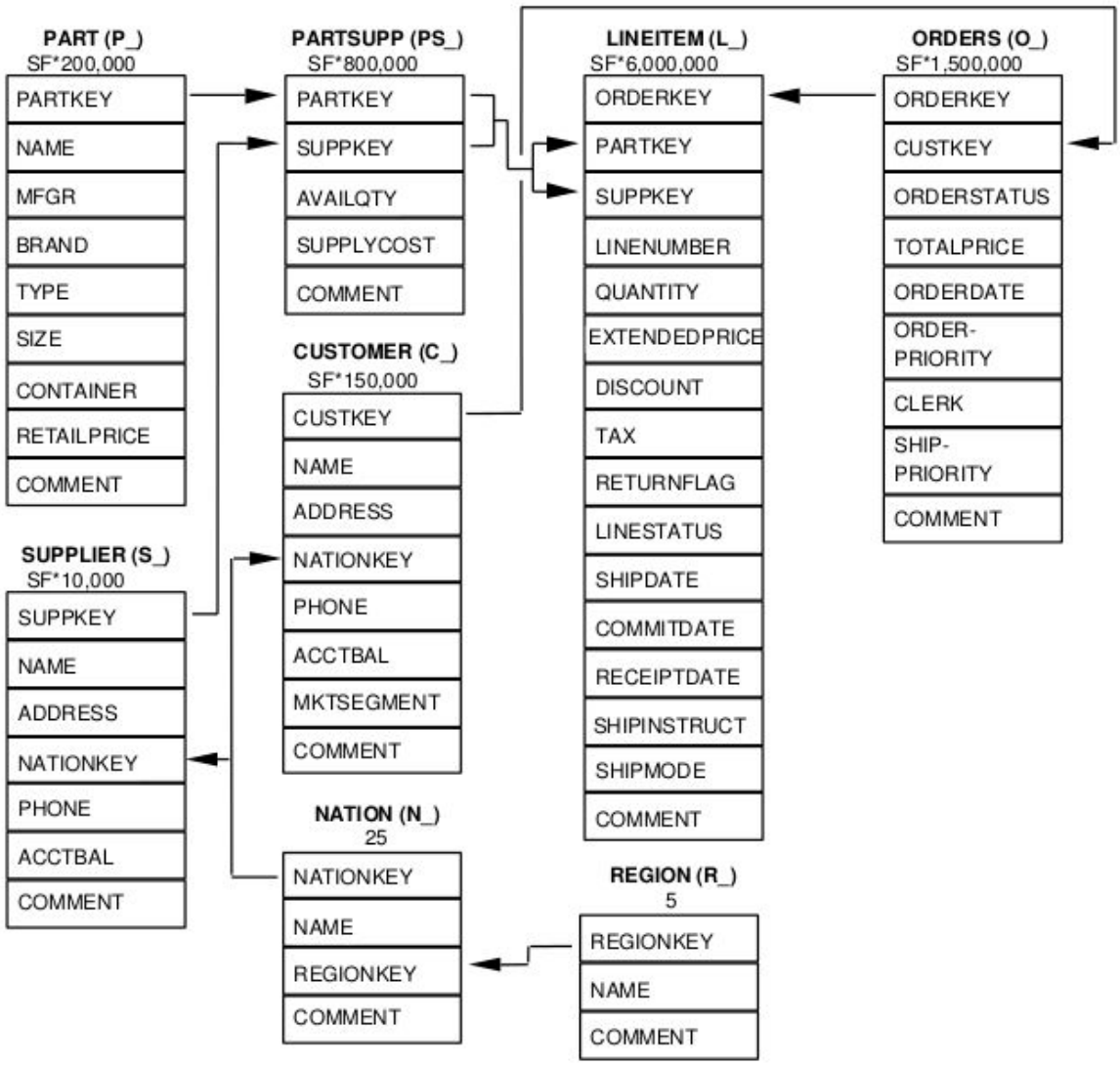
Su grupo como expertos en bases de datos ha sido contratado para el diseño físico de la copia de la base de datos que se va a utilizar para el análisis. Actualmente existe una implementación del modelo lógico con un diseño físico ingenuo donde no se siguió metodología alguna para construirlo. El grupo de expertos debe modificar ese diseño ingenuo con la finalidad de mejorar los tiempos de respuesta y en general el tiempo de ejecución de las consultas propuestas.

Para definir más claramente lo que se quiere hacer con la base de datos, en la sección 2 se presentan las consultas que se realizan con más frecuencia. A diferencia de un ambiente de procesamiento de transacciones en línea (OLTP), en el ambiente de minería de datos se ejecutan consultas "ad-hoc" para obtener resultados y tratar de encontrar patrones, de modo que las consultas presentadas pueden ser ejecutadas en cualquier momento y con cualquier frecuencia, pero todas son importantes para el sistema. La base de datos para minería o "Data Warehouse" es refrescada periódicamente con nuevos datos provenientes de la base de datos de operaciones.

1. Modelo Lógico de FBV

En esta sección se describe el modelo lógico de la base de datos de operaciones de FBV, el cual consiste de ocho (8) relaciones. Para cada relación se especifican los atributos y las restricciones de integridad.

Los nombres de los atributos están en inglés, por compatibilidad con el resto de los sistemas de la empresa que funcionan en coordinación con varias empresas internacionales, las cuales se han puesto de acuerdo en utilizar un lenguaje común en la implementación.



Esquema para la base de datos TPC-H

Relación PART: representa los productos de FBV es decir, bombas (de agua, de achíque, químicas, petroleras, entre otras) y sus partes.

Atributo	Tipo de dato
P_PARTKEY	numeric identifier
P_NAME	variable text, size 55
P_MFGR	fixed text, size 25
P_BRAND	fixed text, size 10
P_TYPE	variable text, size 25
P_SIZE	integer
P_CONTAINER	fixed text, size 10
P_RETAILPRICE	decimal
P_COMMENT	variable text, size 23

Clave Primaria: P PARTKEY

Relación SUPPLIER: representa los fabricantes de insumos de FBV.

Atributo	Tipo de dato
S_SUPPKEY	numeric identifier
S_NAME	fixed text, size 25
S_ADDRESS	variable text, size 40
S_NATIONKEY	numeric identifier
S_PHONE	fixed text, size 15
S_ACCTBAL	decimal
S_COMMENT	variable text, size 101

Primary Key: S_SUPPKEY
S_NATIONKEY Foreign key reference to N_NATIONKEY

Relación CUSTOMER: representa a los clientes que compran productos a través de FBV. El modo de comprar es colocar una orden a través de Internet, donde se especifica cada producto a comprar y en qué cantidad. (Ver relaciones ORDERS y LINEITEM.)

Atributo	Tipo de dato

C_CUSTKEY	numeric identifier
C_NAME	variable text, size 25
C_ADDRESS	variable text, size 40
C_NATIONKEY	numeric identifier
C_PHONE	fixed text, size 15
C_ACCTBAL	decimal
C_COMMENT	variable text, size 117
C_MKTSEGMENT	fixed text, size 10

Primary Key: C_CUSTKEY
C_NATIONKEY Foreign key reference to N_NATIONKEY

Relación PARTSUPP: representa los insumos que son elaborados por cada fabricante.

Atributo	Tipo de dato
PS_PARTKEY	identifier
PS_SUPPKEY	identifier
PS_AVAILQTY	integer
PS_SUPPLYCOST	decimal
PS_COMMENT	variable text, size 199

Primary Key: PS_PARTKEY, PS_SUPPKEY
PS_PARTKEY Foreign key reference to P_PARTKEY
PS_SUPPKEY Foreign key reference to S_SUPPKEY

Relación ORDERS: representa a las órdenes enviadas a FBV por sus clientes.

Atributo	Tipo de dato
O_ORDERKEY	numeric identifier
O_CUSTKEY	numeric identifier

O_ORDERSTATUS	fixed text, size 1
O_TOTALPRICE	decimal
O_ORDERDATE	date
O_ORDERPRIORITY	fixed text, size 15
O_CLERK	fixed text, size 15
O_SHIPPRIORITY	integer
O_COMMENT	variable text, size 79

Primary Key: O_ORDERKEY
O_CUSTKEY Foreign key reference to C_CUSTKEY

Relación LINEITEM: representa los diferentes renglones de cada orden, es decir, las diferentes "líneas" contenidas en una orden.

Atributo	Tipo de dato
L_ORDERKEY	numeric identifier
L_PARTKEY	numeric identifier
L_SUPPKEY	numeric identifier
L_LINENUMBER	integer
L_QUANTITY	decimal
L_EXTENDEDPRICE	decimal
L_DISCOUNT	decimal
L_TAX	decimal
L_RETURNFLAG	fixed text, size 1
L_LINESTATUS	fixed text, size 1
L_SHIPDATE	date
L_COMMITDATE	date
L_RECEIPTDATE	date
L_SHIPINSTRUCT	fixed text, size 25

L_SHIPMODE	fixed text, size 10
L_COMMENT	variable text size 44

L_ORDERKEY Foreign key reference to O_ORDERKEY
L_PARTKEY Foreign key reference to P_PARTKEY
L_PARTKEY,L SUPPKEY Compound Foreign Key Reference to (PS_PARTKEY,PS_SUPPKEY)
Primary Key: L_ORDERKEY, L_LINENUMBER

Relación NATION: representa a los diferentes países con los cuales está asociada la empresa, bien sea porque un fabricante o un cliente están ubicados allí.

Atributo	Tipo de dato
N_NATIONKEY	numeric identifier
N_NAME	fixed text, size 25
N_REGIONKEY	numeric identifier
N_COMMENT	variable text, size 152

Primary Key: N_NATIONKEY
N_REGIONKEY Foreign key reference to R_REGIONKEY

Relación REGION: representa las regiones del mundo en las cuales se encuentran los diferentes paises.

Atributo	Tipo de dato
R_REGIONKEY	numeric identifier 5 regions are populated
R_NAME	fixed text, size 25
R_COMMENT	variable text, size 152

Primary Key: R_REGIONKEY

2. Consultas

En esta sección se especifican algunas consultas representativas y frecuentes que se pueden realizar sobre los datos de FBV. Las consultas tienen nombres de la forma Qij, donde i es un número que corresponde al grupo de clasificación de la consulta y j es el número de la consulta. Primero se da una pequeña explicación de la consulta y luego se coloca su especificación en SQL. Cabe destacar que cada integrante del equipo debe ser el responsable de una consulta en cada grupo de clasificación.

2.1 Q_{11}: Proveedor con el mínimo costo

En esta consulta se encuentra, en una región dada, para cada parte de un cierto tipo y tamaño, el proveedor quien pueda proveerla al costo mínimo. Si distintos proveedores en esa región ofrecen el tipo y tamaño de la parte deseada al mismo costo (mínimo), la consulta lista las partes de los proveedores con los 100 balances de cuenta más altos. Para cada proveedor, la consulta retorna el balance del proveedor, nombre y nación, el número de parte y su fabricante, la dirección del proveedor, número de teléfono y comentarios.

Especificación de Q_{11} en SQL:

```
select s_acctbal,s_name,n_name,p_partkey,      p_mfgr,s_address,s_phone,s_comment
from part, supplier, partsupp, nation, region
where p_partkey = ps_partkey and s_suppkey = ps_suppkey and p_size = $1
    and p_type like $2 and s_nationkey = n_nationkey
    and n_regionkey = r_regionkey
    and r_name = $3 and ps_supplycost = (
        select min(ps_supplycost)
        from partsupp,supplier,nation, region
        where p_partkey = ps_partkey and s_suppkey = ps_suppkey
            and s_nationkey = n_nationkey and n_regionkey = r_regionkey
            and r_name = $3 )
order by s_acctbal desc;
```

datos para probar: $1=15,$2='%BRASS',$3='EUROPE'

2.2 Q_{12}: Prioridad de Envío

En consulta recupera las diez órdenes no enviadas con el valor más alto. La consulta recupera la prioridad de envío y ingreso potencial que se define como la suma de l_extendedprice*(1-l_discount), de las órdenes que tienen los mayores ingresos entre aquellos que no habían sido enviados a partir de una fecha determinada. Los pedidos se enumeran en orden decreciente de los ingresos. Si existen más de 10 pedidos no enviados, sólo se muestran las 10 órdenes con los mayores ingresos.

Especificación de Q_{12} en SQL:

```
select l_orderkey,sum(l_extendedprice*(1 - l_discount)) as revenue,
    o_orderdate,o_shippriority
from customer, orders, lineitem
where c_mktsegment = $1 and c_custkey = o_custkey and l_orderkey = o_orderkey
    and o_orderdate < $2   and l_shipdate > $2
group by l_orderkey, o_orderdate,       o_shippriority
order by revenue desc, o_orderdate;
```

datos para probar: $1='BUILDING',$2='1995-03-15'

2.3 Q_{13}: Reporte de ítems devueltos

Esta consulta identifica los clientes que podrían estar teniendo problemas con las piezas que se les envían. La consulta del reporte de ítems devueltos busca las 20 mejores clientes, en términos de su efecto sobre la pérdida de ingresos para un trimestre dado, que han regresado partes. La consulta considera solamente las piezas que fueron ordenados en el trimestre especificado. La consulta lista el nombre del cliente, dirección, país, número de teléfono, saldo de la cuenta, los comentarios y la pérdida de ingresos. Los clientes se enumeran en orden decreciente de los ingresos

Especificación de Q_{12} en SQL:

```
select c_custkey,c_name,sum(l_extendedprice * (1 - l_discount)) as revenue,
    c_acctbal,n_name,c_address,   c_phone,c_comment
from customer, orders, lineitem, nation
where c_custkey = o_custkey and l_orderkey = o_orderkey
    and o_orderdate >= $1 and o_orderdate < $1 + interval '3 month'
    and l_returnflag = 'R' and c_nationkey = n_nationkey
group by c_custkey, c_name, c_acctbal, c_phone, n_name, c_address, c_comment
order by revenue desc;
```

datos para probar: `$1='1993-10-01'`

2.4 Q_{21}: Modos de envío y orden de prioridad

Esta consulta determina si la selección de los modos menos costosos de envío está afectando negativamente a las órdenes de prioridad crítica debido a que más partes para han sido recibidas por los clientes después de la fecha comprometida. La consulta de modos de envío y de orden de prioridad cuenta por modo de envío, para los *line items* realmente recibidos por los clientes en un año dado, el número de *line items* pertenecientes a las órdenes para las cuales el l_receiptdate excede a l_commitdate para dos modos diferentes de envíos especificados. Sólo *line items* que en realidad fueron enviados antes de la l_commitdate se consideran. Los *line items* finales se dividen en dos grupos, los que tienen prioridad urgente o alto, y los que tienen una prioridad que no sea urgente o ALTO.

Especificación de Q_{12} en SQL:

```
select l_shipmode,sum(case
                          when o_orderpriority = '1-URGENT'
                            or o_orderpriority = '2-HIGH'
                            then 1
                          else 0
                      end) as high_line_count,
                  sum(case
                          when o_orderpriority <> '1-URGENT'
                            and o_orderpriority <> '2-HIGH'
                            then 1
                          else 0
                      end) as low_line_count
from orders,lineitem
where o_orderkey = l_orderkey and l_shipmode in ($1, $2)
    and l_commitdate < l_receiptdate and l_shipdate < l_commitdate
    and l_receiptdate >= $3 and l_receiptdate < $3 + interval '1 year'
group by l_shipmode
order by l_shipmode;
```

datos para probar: `$1='MAIL', $2='SHIP', $3='1994-01-01'`

2.5 Q_{22}: Relación Parte/Proveedor

Esta consulta busca cuántos proveedores pueden suministrar piezas con atributos dados. Se podría utilizar, por ejemplo, para determinar si hay un número suficiente de proveedores de piezas altamente pedidas. Esta consulta cuenta el número de proveedores que puedan suministrar las piezas que satisfagan los requisitos de un cliente en particular. El cliente está interesado en partes de ocho tamaños diferentes, siempre y cuando no sean de un tipo dado, no sean de una determinada marca, y no sean de un proveedor que ha tenido quejas registradas en la Oficina de Buenas Prácticas Comerciales. Los resultados deben ser presentados por número en orden descendente y por marca, tipo y tamaño en orden ascendente.

Especificación de Q_{12} en SQL:

```
select p_brand, p_type, p_size,count(distinct ps_suppkey) as supplier_cnt
from partsupp, part
where p_partkey = ps_partkey and p_brand <> $1
    and p_type not like $2 and p_size in ($3, $4, $5, $6, $7, $8, $9, $10)
    and ps_suppkey not in (
        select s_suppkey
        from supplier
        where s_comment like '%Customer%Complaints%')
group by p_brand,p_type,p_size
order by supplier_cnt desc, p_brand, p_type, p_size;
```

datos para probar: $1='Brand#45', $2='MEDIUM POLISHED%', $3=49, $4=14, $5=23, $6=45, $7=19, $8=3, $9=36, $10=9

2.6 Q_{23}: Oportunidad de Ventas Globales

Esta consulta identifica las geografías en las que hay clientes que pueden ser propensos a realizar una compra. Esta consulta cuenta el número de clientes dentro de un rango específico de códigos de país que no han hecho pedidos por 7 años, pero que tienen un mayor saldo en cuenta que la cuenta promedio "positivo". También refleja la magnitud de la cuenta. El código de país se define como los dos primeros caracteres del c_phone.

Especificación de Q_{12} en SQL:

```
select cntrycode,count(*) as numcust,sum(c_acctbal) as totacctbal
from (
    select substring(c_phone from 1 for 2) as cntrycode, c_acctbal
    from customer
    where substring(c_phone from 1 for 2) in ($1, $2, $3, $4, $5, $6, $7)
        and c_acctbal > (
            select avg(c_acctbal)
            from customer
            where c_acctbal > 0.00
                and substring(c_phone from 1 for 2) in ($1, $2, $3, $4, $5, $6, $7)
           )
        and not exists (select * from orders where o_custkey = c_custkey)
    ) as custsale
group by cntrycode
order by cntrycode;
```

datos para probar: $1='13',$2='31',$3='23',$4='29',$5='30',$6='18',$7='17'

3. Entrega

El proyecto tiene como objetivo el análisis de los requerimientos y de los datos de un caso de estudio y la elaboración de un diseño físico adecuado para la base de datos del caso de estudio. Las actividades específicas a realizar para lograr el objetivo del proyecto son:

1. Estudiar cuidadosamente el dominio del problema, el modelo lógico de la base de datos, además de las consultas más importantes que se desean realizar sobre esos datos.
2. Diseñar organizaciones primarias y caminos de acceso para las relaciones del modelo lógico de FBV, para ello, considere todas las opciones de organizaciones primarias y caminos de acceso vistas en clase de teoría.
3. Debe cargar la base de datos de prueba la cual se puede consultar para obtener estadísticas útiles. el grupo de trabajo debe realizar las siguientes actividades:
   a) Evaluar el volumen de los datos cargados inicialmente. Ud. debe examinar el contenido de la base de datos con consultas apropiadas que le permitan conocer, entre otros aspectos, cuántas tuplas hay en cada tabla, tamaño de cada columna (promedio), hits a shared buffer, hits a cache del sistema operativo y disco duro, cálculo de selectividad, cuántos valores diferentes hay para cada atributo.
   b) Decidir las estadísticas específicas que serían útiles calcular para tomar decisiones de cuáles estructuras de PostgreSQL utilizar para las tablas de la base de datos de FBV. Ud. debe definir por lo menos cuatro (4) estadísticas, además de escribir y ejecutar los comandos en SQL que las calculen.
4. Proponer estructuras de almacenamiento secundario en PostgreSQL (índices si se requieren), para almacenar los datos de las ocho (8) relaciones descritas en el modelo lógico, de modo que las consultas se realicen con un buen rendimiento.

El grupo de trabajo deberá entregar un informe técnico que contenga, además de las secciones habituales, (introducción, conclusiones, etc.), lo siguiente:

- Descripción de las estadísticas. En esta parte del informe se deben documentar las estadísticas utilizadas en el proceso de decisión de las estructuras de almacenamiento. Para ello se debe incluir para cada una: en qué consiste, cómo se calculó la estadística, cuál fue el código fuente utilizado para generarla y los resultados de la ejecución de este código fuente, estos últimos deben ser dispuestos en tablas apropiadas.
- Investigue por lo menos una manera de calcular el tiempo de ejecución de una consulta desde el sistema operativo, es decir no solo calcule el tiempo dentro del manejador (explain analyze).
- Justificación de las estructuras de almacenamiento propuestas. Para cada estructura seleccionada en la actividad descrita en el apartado 4 anterior; justifique su elección, dando argumentos de la utilidad de la estructura y del tipo de operaciones que se van a realizar sobre los datos almacenados en ella, provenientes de las operaciones que se desea realizar.
- *Script* con la creación del modelo físico de la base de datos de FBV. En este *script* se debe incluir la creación de todas las estructuras y parámetros de configuración de PostgreSQL propuestos. En su informe debe justificar los valores utilizados para estos parámetros; si no los especifica o usa los valores por defecto, también debe explicar las razones para hacerlo de esa forma.

- Ejecución del *script* de creación de la base de datos de FBV en su esquema y comprobación de la creación exitosa de todas las estructuras.
- Para el análisis de los costos de ejecución de las consultas, usted deberá tener las dos versiones de la Base de Datos, la original y la propuesta, con los datos cargados. Ejecute las consultas (siempre con cache frío) escogidas en la base de datos original y en la base de datos con las estructuras propuestas, y recoja toda la información necesaria.

  Analice el plan de ejecución de las consultas, dibuje el árbol del plan. De acuerdo a los planes de ejecución, determine si la ejecución de la consulta mejorará con las estructuras propuestas. En caso de que el plan de ejecución no utilice las estructuras propuestas, analice por qué esto sucede, e intente que en el plan se consideren las estructuras propuestas.

  Haga las iteraciones necesarias sobre su(s) estrategia(s) para mejorar las consultas. Para cada iteración debe mostrar los planes de ejecución en los siguientes formatos: i) estándar de PostgreSQL; ii) YAML y iii) un grafo (árbol) que represente el árbol de ejecución propuesto por el *planner*. Debe describir el árbol de ejecución. En este paso, puede explorar la posibilidad de modificar la sintaxis de la consulta de modo que su ejecución sea más eficiente. Analice los datos de la ejecución de la consulta y determine si son consistentes con los planes de ejecución. En cualquier caso, sean los resultados favorables o no para las estructuras propuestas, justifique los resultados obtenidos.

Fecha de Entrega del informe: Miércoles, 15 de junio de 2016, hasta las 3:30 pm

Formato de la entrega: El análisis, la descripción y la justificación se entregan en un informe en papel; el *script* y la muestra de su ejecución se deben enviar por correo electrónico a su profesor de laboratorio colocando en el SUBJECT o ASUNTO “[ci5313] Proyecto GRUPO Gnn" donde nn es el número de grupo.