Universidad Simón Bolívar Septiembre - Diciembre 2014
Arquitectura y Administración de BD (CI-5313)

Construcción de la Base de Datos de *Babel* - Tipo A

Profs. José Tomás Cadenas, Claudia González y Fabiana Reggio

La distribuidora de libros *Babel* ofrece sus servicios a través de Internet a librerías y particulares, para ventas en grandes volúmenes o por libros individuales (ya sea en formato papel o digital). *Babel* es lo que se considera actualmente un "on-line supplier".

Actualmente *Babel* 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.

Babel 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*. El producto principal que vende *Babel* son libros, los cuales se encuentran descritos en la relación *PART*. Además, existe una relación *SUPPLIER* que contiene los proveedores de los libros para los cuales se describe también el país y la región donde se encuentran sus oficinas principales. Cada proveedor puede suministrar muchos libros y un libro puede ser suministrado por muchos proveedores, lo cual se refleja en la relación *PARTSUPP*. Las órdenes de cada cliente se almacenan en la relación *ORDERS* y los detalles (o ítems) de las órdenes se almacenan en la relación *LINEITEM*. Cada línea de detalle de una orden contiene un libro de un proveedor 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 gestor de base de datos seleccionado.

Un grupo de 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 de este modelo lógico con un diseño físico ingenuo y donde no se siguió metodología alguna para construir su modelo interno. La idea es que el grupo de expertos modifique este diseño ingenuo con la finalidad de mejorar los tiempos de respuesta, y en general, el tiempo de ejecución de las consultas propuestas y las actividades de "*refresh*" de la copia de la base de datos. La base de datos existente está instalada con el gestor de base de datos Oracle 11g y tiene una gran cantidad de datos almacenados.

Para definir más claramente lo que se quiere hacer con la base de datos, en las secciones 2 y 3 se presentan las consultas que se realizan con más frecuencia y las operaciones de "*refresh*". 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 Babel

En esta sección se describe el modelo lógico de la base de datos de operaciones de *Babel*, el cual consiste de ocho (8) relaciones. Para cada relación se especifican los atributos y las restricciones de integridad. En el aula virtual del curso se colocó el *script* que permite crear un esquema relacional en Oracle que implementa el modelo lógico presentado. También se crearán las tablas y se cargarán con datos de prueba en un servidor del LDC, para que se puedan estimar los volúmenes de la base de datos.

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.

A continuación se muestra el esquema relacional.

REGION(R_REGIONKEY, R_NAME, R_COMMENT)
NATION(N_NATIONKEY, N_NAME, N_REGIONKEY, N_COMMENT)
N_REGIONKEY REFERENCIA A REGION
PART(P_PARTKEY, P_NAME, P_MFGR, P_BRAND, P_TYPE, P_SIZE, P_CONTAINER, P_RETAILPRICE, P_COMMENT)
SUPPLIER(S_SUPKEY, S_NAME, S_ADDRESS, S_NATIONKEY, S_PHONE, S_ACCTBAL, S_COMMENT)
S_NATIONKEY REFERENCIA A NATION
PARTSUPPLIER(PS_PARTKEY, PS_SUPPKEY, PS_AVAILQTY, PS_SUPPCOST, PS COMMENT)
PS_PARTKEY REFERENCIA A PART
PS_SUPPKEY REFERENCIA A SUPPLIER
CUSTOMER(C_CUSTKEY, C_NAME, C_ADDRESS, C_NATIONKEY, C_PHONE, C_ACCTBAL, C_MKTSEGMENT, C_COMMENT)
C_NATIONKEY REFERENCIA A NATION
ORDERS(O_ORDERKEY, O_CUSTKEY, O_ORDERSTATUS, O_TOTALPRICE, O_ORDERDATE, O_ORDERPRIORITY, O_CLERK, O_SHIPPRIORITY, O_COMMENT)
O_CUSTKEY REFERENCIA A CUSTOMER
LINEITEM (L_ORDERKEY, L_PARTKEY, L_SUPPKEY, L_LINENUMBER, L_QUANTITY, L_EXTENDEDPRICE, L_DISCOUNT, L_TAX, L_RETURNFLAG, L_LINESTATUS, L_SHIPDATE, L_COMMITDATE, L_RECEIPTDATE, L_SHIPINSTRUCT, L_SHIPMODE, L_COMMENT)
(L_PARTKEY, L_SUPPKEY) REFERENCIA A PARTSUPPLIER
L_ORDERKEY REFERENCIA A ORDERS
L_PARTKEY REFERENCIA A PART

Relación PART: representa lo que vende *Babel* es decir, libros.

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

Relación SUPPLIER: representa los proveedores de los libros que vende *Babel*.

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

Relación CUSTOMER: representa a los clientes que compran libros a través de *Babel*. El modo de comprar es colocar una orden a través de Internet, donde se especifica cada libro a comprar y en cuál 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_ MKTSEGMENT	fixed text, size 10
C_ COMMENT	variable text, size 117

Relación PARTSUPP: representa los libros que son suministrados por cada proveedor.

Atributo	Tipo de dato
PS_ PARTKEY	identifier
PS_ SUPPKEY	identifier
PS_ AVAILQTY	integer
PS_ SUPPLYCOST	decimal
PS_ COMMENT	variable text, size 199

Relación ORDERS: representa a las órdenes enviadas a *Babel* 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

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

Relación NATION: representa a los diferentes países con los cuales está asociada la empresa, bien sea porque un proveedor 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

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

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

2. Consultas

En esta sección se especifican las consultas representativas y frecuentes que se realizan sobre los datos de *Babel*. Las consultas tienen nombres de la forma *Qi*, donde *i* es un número de la consulta.

2.1. Q1: Valor de los libros enviados de un país a otro

En esta consulta se determina el valor de todos los libros enviados entre dos países. El valor se refiere a las ganancias derivadas de renglones que corresponden a un libro que se envió de un país al otro especificado, entre los años 1995 y 1996. Se toma en cuenta los renglones donde uno de los países era el cliente y el otro el proveedor y viceversa, pero que sólo involucran a los países identificados en los parámetros de la consulta. El resultado se ordena en forma ascendente de país del proveedor, país del cliente y año.

Especificación de Q1 en SQL:

```
select supp_nation, cust_nation, l_year, sum(volume) as revenue
from (select
n1.n_name as supp_nation,
n2.n_name as cust_nation,
extract(year from l_shipdate) as l_year,
l_extendedprice * (1 - l_discount) as volume
from supplier, lineitem, orders, customer, nation n1, nation n2
where s_suppkey = l_suppkey
and o_orderkey = l_orderkey
and c_custkey = o_custkey
and s_nationkey = n1.n_nationkey
and c_nationkey = n2.n_nationkey
and ((n1.n_name = '&NATION1' and n2.n_name = '&NATION2')
or (n1.n_name = '&NATION2' and n2.n_name = '&NATION1'))
and l_shipdate between date '1995-01-01' and date '1996-12-31')
group by supp_nation,
cust_nation,
l_year
order by supp_nation,
cust_nation,
l_year;
```

2.2. Q2: Prioridad de envío

Esta consulta recupera las diez (10) órdenes no enviadas con el mayor valor para un determinado segmento de mercado (parámetro de entrada). La consulta recupera la prioridad de envío y la ganancia potencial, definida como la suma de l_extendedprice * (1-l_discount), de aquellas órdenes que tienen las mayores ganancias entre aquellas que no han sido embarcadas para una fecha dada (parámetro de entrada). Las órdenes son reportadas en orden descendente de ganancias.

Especificación de Q2 en SQL:

```
select L_ORDERKEY, sum(L_EXTENDEDPRICE*(1-L_DISCOUNT)) as REVENUE,
O_ORDERDATE,
O_SHIPPRIORITY
from CUSTOMER, ORDERS, LINEITEM
where C_MKTSEGMENT = '&segment'
and C_CUSTKEY = O_CUSTKEY
and L_ORDERKEY = O_ORDERKEY
and O_ORDERDATE < '&date'
and L_SHIPDATE > '&date'
group by
L_ORDERKEY, O_ORDERDATE, O_SHIPPRIORITY
order by
REVENUE desc, O_ORDERDATE;
```

2.3. Q3: Relación entre libros y fabricantes

Esta consulta determina cuantos suplidores pueden suplir libros con determinadas características que se refieren a la editorial y tipo de libro. Se quieren totalizar únicamente proveedores que no hayan recibido quejas de sus clientes. Los parámetros de entrada son la editorial y el tipo de libro.

Especificación de Q3 en SQL:

```
select P_BRAND, P_TYPE, count(distinct PS_SUPPKEY) as SUPPLIER_CNT
from PARTSUPPLIER, PART
where P_PARTKEY = PS_PARTKEY and
P_BRAND = '&brand' and
P_TYPE like '&type' and
PS_SUPPKEY not in (select S_SUPPKEY from SUPPLIER
where S_COMMENT like '%Customer%Complaints')
group by
P_BRAND, P_TYPE
order by
SUPPLIER_CNT desc, P_BRAND, P_TYPE;
```

2.4. Q4: Distribución de clientes

Esta consulta determina la distribución de clientes de acuerdo al número de órdenes que han colocado, incluyendo a los clientes que no tienen órdenes.

Especificación de Q4 en SQL:

```
select C_COUNT, count(*) as CUSTDIST
from (select C_CUSTKEY, count(O_ORDERKEY) as C_COUNT
from CUSTOMER left outer join ORDERS on
C_CUSTKEY=O_CUSTKEY
group by C_CUSTKEY)
C_ORDERS
group by
C_COUNT
order by
CUSTDIST desc, C_COUNT desc;
```

2.5. Q5: Proveedor con el menor precio

En esta consulta se consigue cuál proveedor debe ser seleccionado para colocar una orden de un libro de determinado tipo y tamaño en una determinada región. Se desea escoger alguno de los proveedores que lo ofrezca al costo mínimo (pueden haber varios). Los parámetros de entrada son el nombre de la región, el tamaño y el tipo del libro.

Especificación de Q5 en SQL:

```
select PS_SUPPKEY, S_NAME, S_NATIONKEY, S_PHONE, PS_SUPPCOST
from PARTSUPPLIER E, SUPPLIER, NATION, REGION, PART
where PS_PARTKEY=P_PARTKEY and
P_SIZE = &psize and
P_TYPE like '%&ptype' and
R_NAME='&name' and
PS_SUPPKEY = S_SUPPKEY and
S_NATIONKEY = N_NATIONKEY and
N_REGIONKEY=R_REGIONKEY and
PS_SUPPCOST=
(SELECT min( PS_SUPPCOST)
from PARTSUPPLIER I
where E.PS_PARTKEY=I.PS_PARTKEY);
```

3. Operaciones de "*Refresh*"

Una operación de "*refresh*" en este proyecto se refiere a actualizaciones sobre la base de datos de análisis. Nos interesan los dos tipos de operaciones de "*refresh*" que se describen a continuación.

3.1. RF1: Nuevas Ventas

Esta operación le agrega ventas a la base de datos, para ello se insertan nuevas órdenes y renglones de las mismas.

3.2. RF2: Ventas Caducadas

Esta operación elimina ventas a la base de datos, para ello se eliminan algunas órdenes y sus correspondientes renglones de la base de datos.