El Benchmark TPC-H en MySQL y Postgress

Brenda Mariza Quintero Beltrán,

Resumen—Para cualquier Sistema de Información el poder procesar información de manera rápida se ha vuelto una necesidad muy importante. Por este motivo se necesita la construcción de base de datos con características especiales que permitan mejorar los procesos, a esta colección de bases de datos se le denomina Almacenes de Datos (Data Warehouse), éstos favorecen el análisis y divulgación eficiente de los datos (especialmente operaciones de análisis de datos (OLAP).

Palabras Clave—Benchmark, Funciones OLAP, Almacen de Datos, Benchmark TPC-H.

1 INTRODUCCIÓN

A partir de 1999 en el estándar de SQL incluye funcionalidades OLAP, que nos permiten obtener información de bases de datos multidimensionales que son útiles en la toma de decisiones en una empresa u organización.

Para tomar decisiones correctas es necesario implementar un benchmark, el cual puede facilitarnos todas las especificaciones técnicas de un ordenador junto con su rendimiento ante los diferentes estímulos lo que permite realizar comparativas entre diferentes sistemas atendiendo a sus especificaciones técnicas y su rendimiento.

En esta investigación se hará un análisis comparativo en términos de la eficiencia de la funciones OLAP de los Sistemas Gestores de Bases de Datos (SGBD) MySQL y PostgreSQL, para lo cual se tomó como referencia el modelo lógico de la base de datos y las consultas incluidas en el benchmark TPC-H versión 2.8.0, éste se centra como apoyo a la toma de decisiones involucrando grandes cantidades de datos relativamente estables.

• *B.M. Quintero Belrán está con la Escuela de Informática, Universidad Autónoma de Sinaloa, Josefa Ortiz de S/N, Ciudad Universitaria, Culiacán, Sinaloa, 80010, México.*
E-mail: lmarizita@hotmail.com

2 CONCEPTOS BASICOS

Para entender mejor de lo que se trata esta investigación, describiremos lo que son Benchmark, Funciones OLAP, Almacen de Datos y Benchmark TPC-H.

2.1 ¿Que es un Almacen de Datos?

Almacén de Datos (Data Warehouse).- Es una colección de datos en la cual se encuentra integrada la información de la empresa u organización. Esta información es de utilidad en el proceso de toma de decisiones gerenciales.

Un Data Warehouse es como el expediente de una empresa con información transaccional y operacional, que es almacenada en una base de datos diseñada para favorecer análisis y la divulgación eficientes de datos (especialmente OLAP, procesamiento analítico en línea). El almacenamiento de los datos no debe usarse con datos de uso actual.

Los Data Warehouse contienen a menudo grandes cantidades de información que se subdividen a veces en unidades lógicas más pequeñas, llamadas los centros comerciales, dependientes de los datos. Generalmente, dos ideas básicas dirigen la creación de un almacén de los datos:

- Integración de los datos de bases de datos distribuidas con estructuras diferentes,

que facilitan una descripción global y un análisis comprensivo en el almacén de los datos.

- Separación de los datos usados en operaciones diarias de los datos usados en el almacén de los datos para los propósitos de la divulgación, de la ayuda en la toma de decisiones, para el análisis y para controlar.

2.2 Funcionalidades OLAP

OLAP. Es el acrónimo en inglés de procesamiento analítico en línea (On-Line Analytical Processing). Es una solución utilizada en el campo de la Inteligencia de Negocios (Business Intelligence), la cual consiste en consultas a estructuras multidimensionales (o Cubos OLAP) que contienen datos resumidos de grandes Bases de Datos o Sistemas Transaccionales (OLTP). Se usa en informes de negocios de ventas, márketing, informes de dirección, minería de datos y áreas similares.

La razón de usar OLAP para las consultas es la velocidad de respuesta. Una base de datos relacional almacena entidades en tablas discretas si han sido normalizadas. Pero esto es relativamente lento a la hora de realizar las complejas consultas multitabla. Un modelo mejor para búsquedas, es una base de datos multidimensional. La principal característica que potencia a OLAP, es que es lo más rápido a la hora de hacer SELECTS.

2.3 ¿Que es un benchmark?

En términos informáticos un benchmark es una aplicación destinada a medir el rendimiento de un ordenador o de algún elemento del mismo. Para ello se somete a la máquina a una serie de cargas de trabajo o estímulos de distinto tipo con la intención de medir su respuesta ante ellos. De esta forma se puede estimar bajo qué tareas o estímulos un determinado ordenador se comporta de una manera fiable y efectiva o por el contrario se muestra ineficiente.

Esta información es muy útil a la hora de seleccionar una máquina para realizar tareas determinadas en el proceso de postproducción y creación del producto audiovisual, pudiendo elegir la mas adecuada para un proceso determinado. El benchmark también es útil para estimar el nivel de obsolescencia de un sistema o en qué aspectos técnicos puede ser mejorado su rendimiento, por medio de actualizaciones.

Por otro lado el benchmark puede facilitarnos todas las especificaciones técnicas de un ordenador junto con su rendimiento ante los diferentes estímulos lo que permite realizar comparativas entre diferentes sistemas atendiendo a sus especificaciones técnicas y su rendimiento.

Las comparativas son útiles para determinar que características técnicas son las idóneas para conseguir un rendimiento óptimo en una tarea específica. Una comparativa entre múltiples ordenadores de diferentes fabricantes (con diferentes especificaciones técnicas) nos permite determinar a priori cuáles son más adecuados para determinadas aplicaciones y cuáles son mejores para otras.

2.4 Benchmark TPC-H

El Benchmark TPC-H es una prueba de rendimiento a sistemas de soporte de decisiones. Consiste en una ”suite” de negocios orientados a búsquedas convenientes y modificaciones simultáneas de datos. Las búsquedas y la población de datos han sido elegidas un amplia relevancia en la industria. Este benchmark ilustra las decisiones de los sistemas de respaldo que examinan grandes volúmenes de datos, ejecutan búsquedas con un elevado grado de complejidad y responde a situaciones críticas de negocios.

3 EQUIPO Y SOFTWARE UTILIZADO PARA EL EXPERIMENTO

Para llevar a cabo la investigación experimental se utilizó una computadora personal con las siguientes características:

- Procesador: Intel Pentium 4 de 2.66 GHZ.
- Memoria: 512 MB RAM DDR.
- Disco Duro: 160 GB IDE.
- Tarjeta Madre Intel D915GAG/D915PSY.
- El disco duro cuenta con 2 particiones, la partición C: tiene una tamaño de 77.6 GB, y la segunda partición D: tiene un tamaño de 167.3 GB.

La Bases de Datos que se utilizaron para los experimentos fueron creadas en Dbgen, el cual es una herramienta de procesamiento por lotes genérica que se utiliza para generar bases de datos que pueden ser utilizadas en los diferentes Sistemas Gestores de Bases de Datos.

Después de crear la base de datos, éstas fueron migradas a Postgres utilizando el software NAVICAT 8 para PostgreSQL, tambien se migraron las bases de datos a MySQL, utilizan el software NAVICAT 8 para MySQL, los cualos son herramientas que permiten crear, explorar las bases de datos y ejecutar consultas SQL de una manera más fácil.

4 RELACIONES QUE SE CREARON EN MYSQL Y POSTGRESS PARA HACER LAS CCONSULTAS

El modelo lógico de la base de datos del benchmark de TPCH consiste de ocho (8)relaciones. Esta relaciones se crearon en los SGBD MySQL y Postgress, Para cada relación se especifican los atributos y las restricciones de integridad.
A continuación se muestran las referencias a cada una de las relaciones, éstas se encuentran al final del documento:

La Tabla 3 muestra la Relación PRODUCTO, ésta representa lo que se vende.
Clave Primaria: P_NUMERO.

La Tabla 4 muestra la Relación SUPLIDOR, ésta representa a los proveedores de productos.
CLAVE PRIMARIA: S_CVLSUP.
CLAVE FORANEA: S_CVLNACION a N_CVLNACION.

La Tabla 5 muestra la Relación PRODUCTOSUPLIDOR, ésta representa la interrelación entre productos y suplidores.
CLAVE PRIMARIA: PS_NUMERO, PS_CVLSUP.
CLAVE FORANEA: PS_NUMERO A P_NUMERO Y PS_CVLSUP A S_CVLSUP.

La Tabla 6 muestra la Relación CLIENTE, ésta representa a los clientes que compran productos.
CLAVE PRIMARIA: C_CVLCLI.
CLAVE FORANEA: C_CVLNACION A N_CVLNACION A N_CVLNACION.

La Tabla 7 muestra la Relación ORDEN, ésta representa a las órdenes colocadas por los clientes.

CLAVE PRIMARIA: O_CVLORDEN, O_CVLCLI.
CLAVE FORANEA: O_CVLCLI A C_CVLCLI.

La Tabla 8 muestra la Relación RENGLON, la cual representa los diferentes articulos contenidos en una orden, es decir, las diferentes líneas contenidas en una orden.

CLAVE PRIMARIA: L_CVLORDEN, L_NUMRENGLON.
CLAVES FORANEAS: L_CVLORDEN a O_CVLORDEN, L_NUMERO a P_NUMERO y L_CVLSUP a S_CVLSUP.
CLAVE FORANEA COMPUESTA: (L_NUMERO, L_CVLSUP) a (PS_NUMERO, PS_CVLSUP).

La Tabla 9 muestra la Relación NACION, ésta representa a los diferentes países referidos en la base de datos, bien sea porque un suplidor o un cliente est´an ubicados ahi.
CLAVE PRIMARIA: N_CVLNACION.

CLAVE FORANEA: N_CVLREGION a R_CVLREGION.

La Tabla 10 muestra la Relación REGION, ésta representa las diferentes regiones en las cuales se agrupan los países.
CLAVE PRIMARIA: R_CVLREGION.

5 CONSULTAS EN EL BENCHMARK TPC-H

El Benchmark TPC-H contiene una lista de consultas de como apoyo a la toma de deciciones en empresas. En total son 22 consultas, éstas se hicieron en cada uno dde los SGBD(MySQL y PostgreSQL) respectivamente.

A continuación se describe brevemente cada una de ellas.

Consulta Reporte del Resumen de Fijación de Precios (Q1) Esta consulta reporta la cantidad que fuñe facturada, enviada o regresada.

Consulta Minimo Costo Suplidor (Q2) Esta consulta encuentra cual suplidor debería ser seleccionado en orden para un producto dado en una región dada.

Consulta Prioridad de Envío (Q3) Esta consulta regresa las 10 ordenes que no han sido enviadas con el valor mas alto.

Consulta Checar Prioridad Orden (Q4) Esta consulta determina como trabaja el sistema en la prioridad de orden y da una satisfacción al cliente.

Consulta Volumen de un Suplidor Local (Q5) Esta consulta lista del volumen de ingresos por un proveedor local.

Consulta Cambio de Previsión de Ingresos (Q6) Esta consulta cuantifica el total de ingresos que resultan de eliminar cierta compañía.

Consulta Volumen de Envío (Q7) Esta consulta determina el valor de envío entre ciertas naciones para ayudar a renegociar los contratos de envío.

Consulta Cuota de Mercado Nacional (Q8) Esta consulta determina como la cuota del mercado de una nación dada en una región dada, ha cambiado alrededor de 2 años para un tipo de producto.

Consulta Medida de Beneficio por Tipo de Producto (Q9) Esta consulta determina como muchos beneficios provienen de una linea de producto dado, separado por nacion del proveedor y año.

Consulta Reporte Articulo Regresado (Q10) Esta consulta identifica a los clientes quienes han tenido problemas con las productos que les han sido enviados.

Consulta Identificacion de Stock (Q11) Esta consulta encuentra los el stock mas importante de los proveedores en una nacion dada.

Consulta Modo de Envío y Prioridad de Orden (Q12) Esta consulta determina si la selección de los modos de envíos es menos caro está afectando negativamente la prioridad de orden, causando que los clientes reciban productos después de la fecha cometida.

Consulta Destribucion de Clientes (Q13) Esta consulta busca la relación entre clientes y el tamaño de sus ordenes.

Consulta Efecto Promoción (Q14) Esta consulta monitorea la respuesta del mercado a una promoción tal como TV o una campaña especial.

Consulta Top Suplidor (Q15) Esta consulta determina el inicio de manera que pueda ser recompensada, dando mas negocios o identificando un reconocimiento especial.

Consulta Relación Productos/Suplidor (Q16) Esta consulta encuentra como muchos proveedores pueden proveer productos con atributos dados. Esto podría ser utilizado, por ejemplo, para determinar si hay un número suficiente de proveedores para muchas ordenes

de productos.

Consulta Cantidad más pequeña de Ingresos por Orden (Q17) Esta consulta determina como el porcentaje de ingresos por año se perdería si las órdenes no fueran llenadas por pequeñas cantidades de productos. Esto puede sobrepasar los gastos por concentrarse en los en los envíos más grandes.

Consulta Gran Volumen de Clientes (Q18) Esta consulta obtiene un rango de clientes basados en una gran cantidad de órdenes, éstas son definidas por quien tenga una cantidad total de órdenes arriba de cierto nivel.

Consulta Ingresos Descontados (Q19) La consulta de Ingresos Descontados reporta el tamaño de ingresos descontados atribuidos a la venta de artículos seleccionados manejados en una manera particular.

Consulta Promoción de un Producto Potencial (Q20) Esta consulta identifica a los proveedores en una nación particular teniendo productos seleccionados que pueden ser candidatos para una promoción.

Consulta Los proveedores que mantienen órdenes de espera (Q21) Esta consulta identifica ciertos proveedores que no fueron capaces de enviar productos requeridos en un cierto tiempo.

Consulta Oportunidad de Ventas Mundiales (Q22) Esta consulta identifica areas geográficas donde hay clientes quienes pueden hacer una compra.

6 TABLA COMPARATIVA DE RESULTADOS

La tabla 1 muestra los resultados de las comparaciones en cuanto a rendimiento en tiempo de respuesta en cada uno de los SGBD (MySQL y POSTGRESQL).

El experimento se hizo sobre tablas con un tamaño de 100 MB. Los tiempos de respuesta son calculados en segundos para cada consulta.

La tabla 2 muestra los resultados de las comparaciones en cuanto a rendimiento en tiempo de respuesta en cada uno de los SGBD (MySQL y POSTGRESQL).

El experimento se hizo sobre tablas con un tamaño de 1 GB. Los tiempos de respuesta son calculados en segundos para cada consulta.

El experimento no se realizó para BD de 10 GB, ya que no fué posible por falta de tiempo.

7 CONCLUSIONES

Después de ver los resultados que se muestran en la Tabla 1 y Tabla 2, por cada una de las Consultas en cada SGBD (MySQL y PostgreSQL), podemos concluir que, Postgres es menos eficiente que MySQL en consultas en bases de datos pequeñas, por otro lado PostgreSQL es más eficiente que MySQL en BD de mayor tamaño.

REFERENCES

[1] TPC BENCHMARKTM H, *(Decision Support) Standard Specification Revision 2.8.0* LaTeX, 1993 - 2008

[2] Prof. Soraya Abad Mota, *Modelo L´ogico del Benchmark de TPCH* LaTeX, Octubre 2007

TABLE 1
Comparación de Tiempo de Respuesta en BD 100MB

Consulta	Postgress	MySQL
Q1	74.46	17.952
Q2	4.692	235.314
Q3	43.044	23.766
Q4	3.468	2.142
Q5	1.632	30.604
Q6	7.038	6.018
Q7	14.178	6.12
Q8	10.098	1.428
Q9	145.86	10.098
Q10	1.53	1.326
Q11	2.244	1.326
Q12	9.282	6.732
Q13	14.076	16.8
Q14	6.18	31.212
Q15	6.324	20.7
Q16	15.912	8.568
Q17	0.492	1.122
Q18	81	más de 1 hora
Q19	12.954	0.54
Q20	más de 1 hora	0.69
Q21	72	0.05154
Q22	más de 1 hora	0.5

TABLE 2
Comparación de Tiempo de Respuesta en BD 1GB

Consulta	Postgress	MySQL
Q1	744.8	179.62
Q2	46.95	2353.20
Q3	430.54	237.71
Q4	34.74	21.48
Q5	16.48	306.098
Q6	70.42	60.23
Q7	141.88	61.42
Q8	100.99	14.38
Q9	1458.64	100.99
Q10	15.29	13.32
Q11	22.42	13.24
Q12	92.80	67.30
Q13	140.73	168.45
Q14	61.79	312.10
Q15	63.21	207.03
Q16	158.9	85.58
Q17	04.90	11.18
Q18	811.23	más de 1 hora
Q19	12.954	5.4
Q20	más de 1 hora	6.92
Q21	72	0.5154
Q22	más de 1 hora	5.2

TABLE 3
Relación PRODUCTO: representa lo que se vende.

Atributo	Tipo de dato
P_NUMERO	Identificador numérico
P_NOMBRE	Texto variable 55 caracteres
P_MFGR	Texto fijo 25 caracteres
P_MARCA	Texto fijo, 10 caracteres
P_TIPO	Texto variable, 25 caracteress
P_TAMAÑO	entero
P_CONTENEDOR	Texto fijo, 10 caracteres
P_PRECIOVENTA	número decimal
P_COMENTARIO	Texto variable, 23 caracteres

TABLE 4
Relación SUPLIDOR:representa a los proveedores de productos.

Atributo	Tipo de dato
S_CLVSUP	Identificador numérico
S_NOMBRE	Texto fijo, 25 caracteres
S_DIRECCION	Texto variable, 40 caracteres
S_CLVNACION	identificador num´erico
S_TELEFONO	Texto fijo, 15 caracteres
S_SALDOCUENTA	número decimal
S_COMENTARIO	Texto variable, 101 caracteres

TABLE 5
Relación PRODUCTOSUPLIDOR:representa la interrelación entre productos y suplidores.

Atributo	Tipo de dato
PS_NUMERO	Identificador numérico
PS_CLVSUP	identificador numérico
PS_CANTDISP	entero
PS_COSTOSUP	número decimal
PS_COMENTARIO	Texto variable, 199 caracteres

TABLE 6
Relación CLIENTE: Representa a los clientes que compran productos.

Atributo	Tipo de dato
C_CLVCLI	Identificador numérico
C_NOMBRE	Texto fijo, 25 caracteres
C_DIRECCION	Texto variable, 40 caracteres
C_CLVNACION	identificador numérico
C_TELEFONO	Texto fijo, 15 caracteres
C_SALDOCUENTA	número decimal
C_SEGMERC	Texto fijo, 10 caracteres
C_COMENTARIO	Texto variable, 101 caracteres

TABLE 7
Relación ORDEN: representa a las órdenes colocadas por los clientes.

Atributo	Tipo de dato
O_CLVORDEN	Identificador numérico
O_CLVCLI	Identificador numérico
O_STATUSORDEN	Texto fijo, 1 caracter
O_PRECIOTOTAL	número decimal
O_FECHAORDEN	fecha
O_PRIORIDADORDEN	Texto fijo, 15 caracteres
O_EMPLEADO	Texto fijo, 15 caracteres
O_PRIORIDADENVIO	entero
O_COMENTARIO	Texto variable, 79 caracteres

TABLE 8
Relación RENGLON, ésta representa los diferentes items contenidos en una orden, es decir, las diferentes líneas contenidas en una orden.

Atributo	Tipo de dato
L_CLVORDEN	Identificador numérico
L_NUMERO	Identificador numérico
L_CLVSUP	identificador num´erico
L_NUMRENGLON	entero
L_CANTIDAD	número decimal
L_PRECIOEXT	número decimal
L_DESCUENTO	número decimal
L_IMPUESTO	número decimal
L_RETURNFLAG	Texto fijo, 1 caracter
L_STATUSRENGLON	Texto fijo, 1 caracter
L_FECHAENVIO	fecha
L_FECHACOMMIT	fecha
L_FECHARECIBIDO	fecha
L_ENVIOINSTR	Texto fijo, 25 caracteres
L_ENVIOMODO	Texto fijo, 10 caracteres
L_COMENTARIO	Texto variable, 44 caracteres

TABLE 9
Relación NACION: representa a los diferentes países referidos en la base de datos, bien sea porque un suplidor o un cliente están ubicados allí.

Atributo	Tipo de dato
N_CLVNACION	Identificador numérico
N_NOMBRE	Texto fijo, 25 caracteres
N_CLVREGION	identificador numérico
N_COMENTARIO	Texto variable, 152 caracteres

TABLE 10
Relación REGION: representa las diferentes regiones en las cuales se agrupan los países.

Atributo	Tipo de dato
R_CLVREGION	Identificador numérico
R_NOMBRE	Texto fijo, 25 caracteres
R_COMENTARIO	Texto variable, 152 caracteres