**MULTICIENCIAS**, Vol. 11, Nº 4, 2011 (410 - 415)
ISSN 1317-2255 / Dep. legal pp. 200002FA828

# Rendimiento de consultas SQLf en arquitecturas débil y fuertemente acopladas

**José Tomás Cadenas[1,2], Ana Aguilera[2] y Leonid Tineo[1,2]**

[1] *Departamento de Computación y T.I., Universidad Simón Bolívar. Caracas, Venezuela.*
[2] *Centro de Análisis, Modelado y Tratamiento de Datos, FACYT, Universidad de Carabobo. Valencia, Venezuela*

jtcadenas@usb.ve , aaguilef@uc.edu.ve , leonid@ldc.usb.ve

## Resumen

En este artículo se presenta una comparación del rendimiento de consultas difusas implementadas bajo diversas estrategias: fuertemente acoplado (PostgreSQLf) y débilmente acoplado (SQLfi) versus consultas clásicas (precisas). Uno de los principales problemas en la comunidad de base de datos para adoptar herramientas que ofrecen mayores beneficios es la resistencia de añadir costos y tiempo de procesamiento que degrade el rendimiento de los sistemas de *software*. Los sistemas gestores de bases de datos (DBMS) en sí son complejos, por lo cual se aceptan extensiones si tienen un buen rendimiento. Se propone una comparación para medir el rendimiento de cada estrategia de implementación. Se diseñaron experimentos computacionales basados en modelos estadísticos para verificar la validez. Se muestra que el tiempo de ejecución de las consultas con SQLfi es mucho mayor que el de PostgreSQLf.

**Palabras clave:** procesamiento de consultas difusas, PostgreSQL, SQLf, DBMS.

Recibido: 21-11-2010 / Aceptado: 08-12-2011

# Performance of SQLf Queries about Loose and Tight Coupling in Architecture


## Abstract

This article presents a comparison of the performance of the fuzzy query system implemented for various strategies, tight coupling (PostgreSQLf) and loose coupling (SQLfi), versus classical (crisp) queries. One of the major problems of database community resistance to adopting tools that offer greater benefits is the reluctance to adding costs and processing time that degrade software system performance. Database management systems (DBMS) are complex in themselves, so they accept extensions if these can demonstrate good performance. This study proposes a comparison to measure the efficiency of each implementation strategy. Computational experiments were designed based on statistical models for verifying validity. It is shown that the runtime for queries with SQLfi is much higher than with PostgreSQLf.


**Keywords:** fuzzy query processing, PostgreSQL, SQLf, DBMS.

## 1. Introducción

Se observa un gran interés en la comunidad científica de bases de datos en el manejo de datos inciertos; de esta forma, MystiQ desarrollado en la Universidad de Washington por Boulos et al (2005) es un sistema que usa un modelo de datos probabilístico para encontrar respuestas en un gran número de fuentes de datos exhibiendo varios tipos de imprecisiones. Trio, concebido en la Universidad de Stanford por Widom (2009), es un Sistema Gestor de Base de Datos donde el dato, la incertidumbre y el linaje (procedencia del dato) son gestionados en forma integrada. Koch (2009) propone MayBMS, un sistema para la administración escalable de información incierta (utilizando un modelo probabilístico) implementado sobre PostgreSQL en la Universidad de Cornell. Desde hace varios años investigadores han modelado la incertidumbre en los sistemas de gestión de base de datos utilizando la lógica difusa Zadeh (1965) y la teoría de conjuntos difusos de Zadeh (1995); un compendio de trabajos recientes lo presenta Galindo (2008).

Connolly (2005) plantea para los usuarios el buen rendimiento de los sistemas gestores de bases de datos (DBMS) es una condición necesaria para ser aceptados. Sin embargo el problema ha sido relegado a un segundo plano en las investigaciones sobre sistemas gestores de bases de datos difusas tal como lo presenta López (2006). En este artículo se comprueba lo planteado por Aguilera (2011) que una estrategia de implementación fuertemente acoplada tiene un mejor rendimiento que una débil para extender las prestaciones de un DBMS que admita utilizar SQLf con aval en Bosc (1995).

Bosc (2000) refiere que SQLf es un lenguaje de consultas difusas diseñado sobre DBMS relacionales, permitiendo según Urrutia (2008) el uso de una condición difusa en cualquier lugar donde se permite una condición booleana clásica. Tales condiciones involucran etiquetas lingüísticas (Zadeh, 1989) con una semántica definida por el usuario que expresa sus preferencias. Para una explicación detallada de sintaxis y aplicaciones desarrolladas utilizando SQLf y su evolución, se puede revisar Cadenas (2010), Goncalves (2008), Goncalves (2001a) y Goncalves (2001b). Por su parte Timarán (2001) asume que existen diversas estrategias para integrar extensiones en los DBMS: acoplamiento fuerte, medio o débil. En este artículo se compara el rendimiento de consultas clásicas, acoplamiento débil "SQLfi" de Goncalves, (2008) y fuerte PostgreSQLf por Aguilera (2009) de consultas difusas.

## 2. Metodología

Se va a comprobar la existencia de diferencia significativa estadísticamente (Berman, 2007) en los tiempos de respuesta de consultas difusas a base de datos y consultas

clásicas. Para las consultas difusas se permite la utilización de términos difusos en la condición de la consulta (predicados difusos que pueden ser unimodal o monótonos), obteniendo como resultado filas con grado de acuerdo a la función de pertenencia predefinida en cada caso. Por otro lado se hicieron consultas clásicas sobre la base de datos utilizando el DBMS PostgreSQL versión 8.3, tomando en cuenta que estas consultas corresponden a las que derivan las consultas difusas efectuadas sobre PostgreSQLf y SQLfi, para ello se utiliza el principio de derivación propuesto por Bosc y Pivert (Bosc, 2000).

Se pobló una base de datos PostgreSQL versión 8.3 utilizando el TPC Benchmark versión 2.9.0 (TPC), generando dos volúmenes de datos diferentes, factor de escala 1 (equivale a 1 GB) y factor de escala 10 (equivale a 10 GB). A estos volúmenes se les denominó bajo y alto respectivamente. Se utilizó el mismo computador para todos los experimentos, Compaq® Presario con procesador Intel® core duo 1.83 GHz, 3 Gbyte en memoria RAM y disco duro de 160 Gbyte, sistema operativo Ubuntu 9.04. Además al tomar el tiempo de respuesta de la consulta se inició el servidor de PostgreSQL para mantener las condiciones iniciales. Luego se crearon consultas variando diversos factores en dos niveles, a saber: Número de Tablas (1 y 3), Número de predicados difusos unimodales (1 y 2), Número de predicados difusos monótonos (1 y 2). La combinación de cada uno de estos factores da como resultado ocho (8) consultas, las cuales se realizaron para cada volumen de datos (16 consultas), cada consulta se repitió una vez con los mismos factores (32 consultas), que luego se hicieron para cada tipo de acoplamiento (débil, fuerte y ninguno), por lo que se realizaron 96 experimentos en los cuales se midió el tiempo de respuesta (tiempo total de ejecución) y el número de filas resultado. Se ha demostrado en otros estudios entre los que destacan Cadenas (2008) y Tineo (2006) con los prototipos de acoplamiento débil y fuerte que estos factores tienen influencia en los tiempos de respuesta. Mediante el número de filas resultado se comprueba funcionalmente que se está obteniendo el mismo resultado de cada una de las consultas (32) sobre los diferentes tipos de acoplamiento.

## 3. Resultados

En la Tabla 1 se resumen estadísticos descriptivos obtenidos por SPSS (SPSS, 2006), se muestra una gran diferencia entre la media y la mediana (4.9450 segundos), esta última no está influenciada por valores atípicos (*outliers*).

Tabla 1. Estadísticos descriptivos.

| N | Válidos | 96 |
|---|---|---|
| | Perdidos | 0 |
| Media | | 276.4136 |
| Mediana | | 4.9450 |
| Moda | | .83(a) |
| Desv. típ. | | 1003.55607 |
| Mínimo | | .61 |
| Máximo | | 6553.94 |
| Percentiles | 25 | 1.3600 |
| | 50 | 4.9450 |
| | 75 | 18.6000 |

Existen varias modas. Se muestra el menor de los valores.
Fuente: autores.

En cuanto a la dispersión, se tiene que la diferencia entre el valor mínimo y el máximo es alta (0.61 y 6553.94 segundos), además debido a que los dos primeros cuartiles están cerca (1.36 y 4.945 segundos) significa que un 25% de la muestra tiene tiempos de respuesta muy parecidos, por lo que existe una concentración en ese intervalo. Se puede observar gráficamente la dispersión en el diagrama de caja presentado en la Figura 1, se utilizó una escala logarítmica con base 10 para el tiempo, se tomó como etiqueta de valores atípicos el número de tablas; se nota que cuando existen 3 tablas se producen los valores atípicos. La mayor concentración de tiempos de respuesta de consultas entre 0,61 y 4.9450, donde se ubican el 50% de las consultas, más aún hasta 18,6 segundos está el 75% de las mismas, observándose gran dispersión en el cuarto cuartil, con varios puntos *outliers* después de los 1000 segundos.

El resultado al utilizar como variable de agrupación el tipo de acoplamiento se visualiza en la Figura 2. Se observa que los mayores tiempos de respuesta se presentan con el tipo de acoplamiento débil, además son los que influyen en la variabilidad.

En cuanto a los tiempos de respuesta del acoplamiento fuerte y ninguno están entre el mínimo y 20 segundos. El resumen de las medias agrupadas por tipo de acoplamiento puede observarse en la Tabla 2.

Se observa que para el caso de acoplamiento débil, tanto la media y la desviación típica son altas con respecto a los otros dos tipos de acoplamiento. Para comprobar estadísticamente la diferencia significativa de las medias entre los tipos de acoplamiento débil y los otros dos, se realizó un test estadístico ANOVA (Tabla 3).

Se puede asegurar que existe por lo menos una diferencia entre las medias de las agrupaciones por tipo de acoplamiento, tal como se visualiza en la Figura 3, donde se nota la drástica diferencia entre las medias del acoplamiento fuerte y ninguno con respecto al tipo de acoplamiento débil.

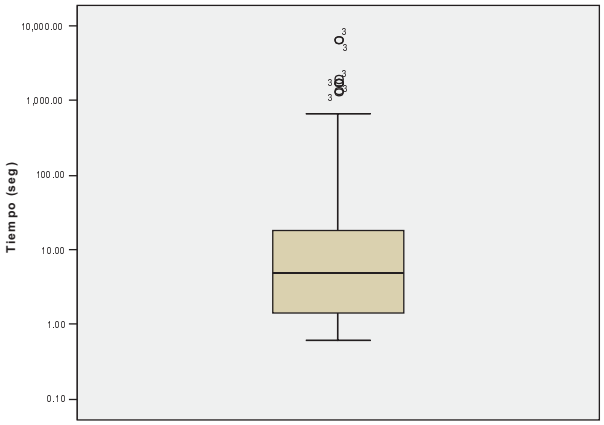
Mediante un análisis post-hoc de Tukey, se reafirma lo observado y no se pueden establecer diferencias con las muestras de datos entre los tipos de acoplamiento Fuerte y Ninguno, de acuerdo a lo mostrado en la Tabla 4.

## 4. Consideraciones finales

A través del análisis estadístico formal, se pudo comprobar que existe diferencia entre los tiempos de respuesta de los tipos de acoplamiento (fuerte, débil y ninguno) de los sistemas de consultas difusas, donde no hay diferencias significativas entre las medias muestrales del acoplamiento fuerte y ninguno. Esto es un resultado halagador para la implementación del tipo de acoplamiento fuerte, ya que indica que se están ofreciendo beneficios novedosos (consultas difusas) en sistemas gestores de bases de datos sin afectar significativamente el rendimiento. Se plantea como retos futuros seguir comparando el comportamiento del acoplamiento fuerte versus las consultas clásicas tomando en cuenta otros tipos de términos difusos en la consulta: comparadores, conectores, modificadores, cuantificadores difusos y predicados (por extensión y expresión).

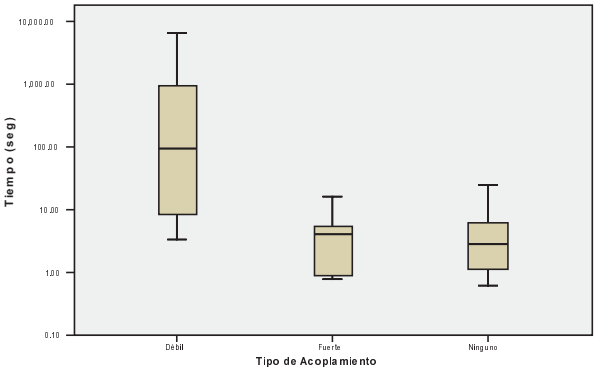
## Agradecimientos

Al FONACIT (Proyecto No. G-2005000278) por su apoyo para el desarrollo de esta investigación. Dad gracias al Señor porque es bueno, porque es eterna su misericordia (Salmo 117).




Fuente: autores

**Figura 1.** Diagrama de Caja (escala logarítmica).




Fuente: autores

**Figura 2.** Diagrama de Caja agrupado por Tipo de Acoplamiento.

Tabla 2. Medias por tipo de Acoplamiento (tiempo en segundos).

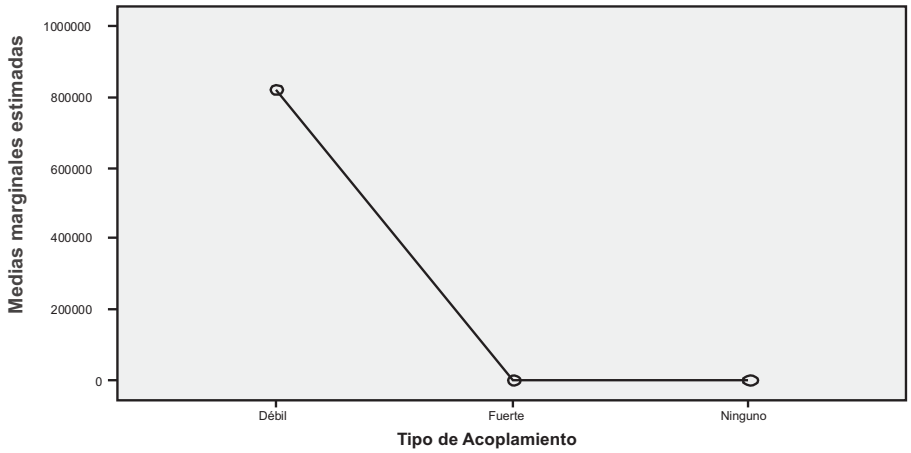
| Tipo de Acoplamiento | Media | N | Desv. típ. |
|---|---|---|---|
| Débil | 819.8222 | 32 | 1621.44022 |
| Fuerte | 4.8188 | 32 | 4.78318 |
| Ninguno | 4.6000 | 32 | 5.58951 |
| Total | 276.4136 | 96 | 1003.55607 |

Fuente: autores.

Tabla 3. Pruebas de los efectos inter-sujetos.

| Fuente | Suma de cuadrados tipo III | Gl | Media cuadrática | F | Significación |
|---|---|---|---|---|---|
| Modelo corregido | 14174056299282.950(a) | 2 | 7087028149641.470 | 8.087 | .001 |
| Intersección | 7334791846357.370 | 1 | 7334791846357.370 | 8.369 | .005 |
| Acoplamiento | 14174056299282.980 | 2 | 7087028149641.490 | 8.087 | .001 |
| Error | 81502785627573.000 | 93 | 876374039006.162 | | |
| Total | 103011633773213.400 | 96 | | | |
| Total corregida | 95676841926856.000 | 95 | | | |

Variable dependiente: Tiempo (ms), R cuadrado =.148 (R cuadrado corregida =.130).
Fuente: autores.




Fuente: autores

**Figura 3.** Gráfico Medias por Acoplamiento.

Tabla 4. Subconjuntos Homogéneos (Tiempo ms).

| Tipo de Acoplamiento | N | Subconjunto | |
|---|---|---|---|
| | | 2 | 1 |
| Ninguno | 32 | 4598.66 | |
| Fuerte | 32 | 4818.58 | |
| Débil | 32 | | 819821.41 |
| Significación | | 1.000 | 1.000 |

Fuente: autores.

## Referencias


AGUILERA, Ana; CADENAS, José; TINEO, Leonid (2011). Fuzzy Querying Capability at Core of an RDBMS. In Yan, L., & Ma, Z.(Eds.) Advanced Database Query Systems: Techniques, Applications and Technologies. **IGI Global**. New York, EEUU, pp. 160-184.

AGUILERA, Ana; CADENAS, José; TINEO, Leonid (2009). Desarrollo de SQLf en Capa Física. **Tecnología, Gerencia y Educación**. Vol. 10, pp. 25-37.

BERMAN, R.; SAUNDERS, M. (2007). **Dealing with Statistics. What you Need to Know**. **Open University Press. Buckingham, GBR.** McGraw-Hill Education.

BOSC, P.; PIVERT, O. (1995). SQLf: A Relational Database Language for Fuzzy Quering. **IEEE Transactions on Fuzzy Systems**, Vol. 3, No. 1.

BOSC, P.; PIVERT, O. (2000). SQLf query functionality on top of a regular relational DBMS. In Pons, Vila, and Kacprzyk (Eds.) **Knowledge Management in Fuzzy Databases,**

BOULOS, J.; DALVI, N.; MANDHANI, B.; MATHUR, S.K; RE, C.; SUCIU, D. (2005) MystiQ: A system for finding more answers by using probabilities. System Demo in ACM SIGMOD International Conference on Management of Data.

CADENAS, J.T. (2008). Una Contribución a la Interrogación Flexible de Bases de Datos: Optimización y Evaluación a Nivel Físico. Trabajo de Grado de Maestría presentado en la Universidad Simón Bolívar, Venezuela.

CODD, E.F. (1970). A relational model of data for large Shared data Banks. **Communications of the ACM**, 13 (6):377-387

CONNOLLY, T.; BEGG, C. (2005). **Database Systems - a Practical Approach to Design, Implementation and Management.** Pearson Education Limited, United Kingdom.

GALINDO, J. (Ed.), (2008). **Handbook of Research on Fuzzy Information Processing in Databases.** Hershey, PA, USA: Information Science.

GONCALVES, M.; TINEO, L. (2008). SQLfi and its Applications. **Avances en Sistemas e Informática**, Vol 5 No. 2. Medellin, ISSN 1657-7663.

GONCALVES M., TINEO L. (2001a). SQLf Flexible Querying Language Extension by means of the norm SQL2, In Proc of FUZZ-IEEE.

GONCALVES, M.; TINEO, L. (2001b). SQLf3: An extension of SQLf with SQL3 features, In Proc of FUZZ-IEEE.

KOCH, C. (2009). MayBMS: A Database Management System for Uncertain and Probabiilistic Data. In Aggarwal (Ed.), **Managing and Mining Uncertain Data** pp.149-184.

LÓPEZ, Y.; TINEO, L. (2006). About the Performance of SQLf Evaluation Mechanisms. **CLEI Electronic Journal**. Volume 9, 2, Paper 8.

SPSS 15.0 (2006) para Windows. Version 15.0.1. Copyrigth © SPSS Inc., 1989-2006.

TIMARÁN, R. (2001). Arquitecturas de Integración del Proceso de Descubrimiento de Conocimiento con Sistemas de Gestión de Bases de Datos: un Estado del Arte, Ingeniería y Competitividad, 3(2).

TINEO, L. (2006). A Contribution to Database Flexible Querying: Fuzzy Quantified Queries Evaluation. Disertación doctoral, Universidad Simón Bolívar, Venezuela.

TPC. Transaction Processing Performance Council. Disponible en: http://www.tpc.org/ [Consultado 05/02 /2011]

URRUTIA, A., TINEO, L, GONZÁLEZ, C. (2008).FSQL and SQLf: Towards a Standard in Fuzzy Databases. In Galindo

(Ed.) **Handbook of Reasearch on Fuzzy Information Processing in Databases**, Volume 1, pp 270-298. Idea Group Inc.

WIDOM, J. (2009). Trio: A system for Integrated Management of Data, Uncertainty, and Lineage. In Aggarwal (Ed.), **Managing and Mining Uncertain Data** (pp. 113-148).

ZADEH, L.A. (1965). Fuzzy Sets, **Information and Control,** 8:338-353.

ZADEH, L. (1989). Knowledge Representation in Fuzzy Logic. **IEEE Transactions on Knowledge and Data Engineering**. Vol 1, No. 1.

ZADEH, L. (1994) Soft Computing and Fuzzy Logic. **IEEE Software** 11 (6), 48-56.