

# Optimización de Esquemas de Consulta para un modelo de Acceso Remoto a Datos

## AUTORES

LEONCIO JIMÉNEZ CANDIA  
Departamento de Computación e Informática  
Universidad Católica del Maule, Talca – Chile  
[Ljimenez@spock.ucm.cl](mailto:Ljimenez@spock.ucm.cl)

JORGE IBÁÑEZ ESPINOSA  
Ingeniero de Ejecución en Computación e Informática  
Universidad Católica del Maule  
[jibaneze@pf.cl](mailto:jibaneze@pf.cl)

FERNANDO SAN MARTIN WOENER  
Empresa E-Linx  
[Snmartin@e-linx.cl](mailto:Snmartin@e-linx.cl)

## RESUMEN

En este artículo se presentan los resultados [4] de un análisis estadístico no paramétrico, aplicando la Prueba T Wilcoxon para comparar las diferencias de tiempos de respuesta de dos Sistemas de Gestión de Bases de Datos, que pertenecen a una empresa del rubro de la construcción. En particular se han comparado, por una parte, los tiempos de respuesta de esquemas anidados de consulta ejecutados en *PostgreSQL*<sup>1</sup>, y por otra parte, los tiempos de respuesta de esquemas iterados de consulta ejecutados en *MySQL*<sup>2</sup> y *PostgreSQL*<sup>3</sup>. Esta comparación nos parece interesante, ya que por ejemplo, en la obtención de reportes mediante una aplicación Cliente/Servidor, los tiempos de respuestas –varían– según el tipo de consulta y plataforma usada. Por otra parte, los resultados obtenidos y que son discutidos en este artículo, corresponden a un modelo de acceso remoto a datos real de una empresa en particular. Por consiguiente, esto no nos permite generalizarlos, sin embargo, el hecho de utilizar un análisis estadístico no paramétrico nos parece un buen punto de reflexión para seguir en esa dirección.

*Palabras Clave:* SGBD, MySQL, PostgreSQL, Esquema de Consulta, Análisis estadístico no paramétrico, Prueba T Wilcoxon.

---

<sup>1</sup> Versión 7.2.

<sup>2</sup> Versión 3.23.46.

<sup>3</sup> Versión 7.2.

## 1. Introducción

El explosivo crecimiento en los volúmenes de los datos implica un replanteamiento permanente de las características funcionales de los SGBD (Sistema de Gestión de Bases de Datos) tales como: la gestión distribuida de los datos, su fiabilidad y normalización [1,2,5] los lenguajes de consultas utilizados, sus tiempos de respuesta y consumación de memoria [4] el control de la concurrencia [6] etc. Esta preocupación, no solamente es válida para empresas desarrolladoras de bases de datos comerciales, como Oracle o SQLServer, pero sino también para SGBD de libre distribución, tales como: MySQL y PostgreSQL. En este sentido, es bueno tener presente que estos motores son los de mayor uso por las pequeñas y medianas empresas de hoy en día, dado que son de dominio público.

En este artículo se ofrece un análisis estadístico que permite comparar, por una parte, las diferencias de tiempos de respuesta de esquemas anidados de consulta ejecutados en *PostgreSQL*, y por otra parte, las diferencias de tiempos de respuesta de esquemas iterados de consulta ejecutados en *MySQL* y *PostgreSQL*. En este sentido, dada la naturaleza de las pruebas, en que se desconoce la distribución de los datos, por una parte y por otra parte, del hecho que no es posible estimar los parámetros poblacionales (media, varianza, etc.) a partir de una muestra aleatoria, fue necesario el uso de un método estadístico no paramétrico. En esta investigación se utilizó el método propuesto por Wilcoxon y Wilcox [9] llamado: *la prueba de Wilcoxon para intervalos con signo*, o simplemente: *T Wilcoxon*. Esta elección se debe al hecho de que los otros métodos no paramétricos, consideran solamente las diferencias entre las parejas de observaciones y no toman en cuenta el orden por rango de sus diferencias en valor absoluto.

Todas las pruebas T Wilcoxon para diferencias de tiempos de respuesta fueron realizadas para bases de datos equivalentes respecto al modelo de acceso remoto a datos, en tercera forma normal (3FN) [8] y la cantidad de tuplas fue mantenida constante para cada una de las tablas utilizadas. Es importante destacar que la experiencia fue realizada en un modelo de acceso remoto a datos real y de propiedad de una conocida empresa constructora de la Región del Maule.

El trabajo se inicia, entonces, con la presentación de los esquemas de consulta y se define lo que se entiende por *tiempo de respuesta*. En cada caso los esquemas de consulta son ilustrados con las consultas hechas al SGBD y que corresponden al modelo de acceso remoto a datos de la empresa constructora. Enseguida, se introduce el marco teórico de la prueba T Wilcoxon y su contexto de aplicación. Luego, se describen las pruebas realizadas y se comentan sus resultados. Finalmente, se presentan las conclusiones y trabajos futuros.

## 2. Esquemas de consulta

El concepto *esquema de consulta* [7] se refiere a la consulta que se realiza al SGBD bajo el modelo de acceso remoto a datos en que el cliente solicita un conjunto de tuplas al SGBD. En este trabajo se han utilizado dos tipos. La figura 1 muestra el esquema anidado de consulta, mientras que en la figura 2 muestra el esquema iterado de consulta.

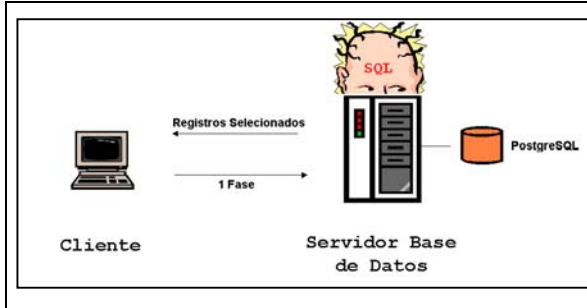


Figura 1. Esquema anidado de consulta

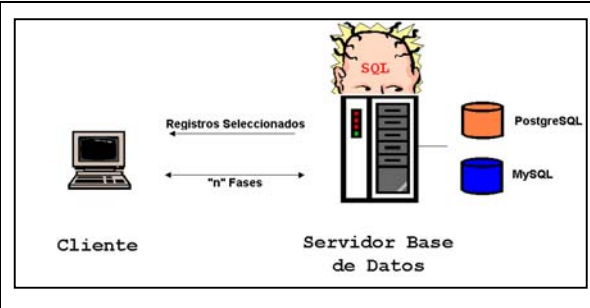


Figura 2. Esquema iterado de consulta

En que:

- a) Esquema anidado de consulta. Este esquema envía una cadena SQL al SGBD en *una fase*. Es decir, las sentencias SQL se ejecutan en una anidación de sentencias SQL en el servidor PostgreSQL. En cambio las anidaciones en el cliente PostgreSQL se consiguen a partir de un ciclo while(rs.next()) escrito en un programa en lenguaje Java, que se puede consultar en [4], para cada uno de los esquemas anidados de consulta considerados: esquema de consulta Not Exist<sup>4</sup> y esquema de consulta Not In<sup>5</sup>.
- b) Esquema iterado de consulta. Este esquema envía una cadena SQL al SGBD en *más de una fase*. Es decir, las consultas son divididas en el servidor (PostgreSQL o MySQL) en consultas parciales o subconsultas individuales, y luego unidas mediante iteraciones de objetos de datos ResultSet en lenguaje Java en el cliente (PostgreSQL o MySQL). Este programa se puede consultar en [4]. Los esquemas considerados fueron: esquema de consulta simple<sup>6</sup>, esquema de consulta condicional<sup>7</sup>, esquema de consulta Inner Join<sup>8</sup> y esquema de consulta condicional compleja<sup>9</sup>.

En ambos casos, el *tiempo de respuesta* se define como el tiempo (en segundos) transcurrido desde que se envía el esquema de consulta implementado, como una cadena SQL, hasta que el SGBD en cuestión, envía las tuplas seleccionadas de vuelta al cliente (en caso de haber alguna), registrándose la cota superior del intervalo de tiempo. Para obtener los datos, fue implementada una aplicación en el cliente en lenguaje Java, conectada al SGBD con una interfaz JDBC (Java Data Base Connectivity). Se trata de un proceso batch que se ejecuta en el front-end y que entrega los tiempos de respuesta. Este programas se ejecuta tantas veces como lo indique el tamaño muestral de la prueba. Respecto a los SGBD utilizados, estos fueron instalados en un mismo servidor de base de datos, para evitar diferencias de configuración de hardware (velocidad de discos, áreas de buffer, etc.). Los detalles de la instalación se encuentran en [4].

### 3. Prueba de T Wilcoxon

Esta prueba fue introducida por Wilcoxon y Wilcox [9]. En que dada una muestra aleatoria de una población con distribución continua y simétrica, la prueba de T Wilcoxon se usa para

<sup>4</sup> SELECT a.cod\_lote FROM lote a WHERE NOT EXISTS (SELECT b.cod\_lote FROM contrato b WHERE b.cod\_etapa = 31 AND b.cod\_proyecto = 14 AND a.cod\_lote = b.cod\_lote) AND a.cod\_etapa = 31 AND a.cod\_proyecto = 14.

<sup>5</sup> SELECT cod\_lote FROM lote WHERE cod\_lote NOT IN (SELECT cod\_lote FROM contrato WHERE cod\_etapa = 31 AND cod\_proyecto = 14) AND cod\_etapa = 31 AND cod\_proyecto = 14.

<sup>6</sup> SELECT DISTINCT rut\_cliente FROM contrato.

<sup>7</sup> SELECT DISTINCT rut\_cliente, proyecto.nom\_proyecto FROM contrato, proyecto WHERE contrato.cod\_proyecto = proyecto.cod\_proyecto.

<sup>8</sup> SELECT DISTINCT rut\_cliente, proyecto.nom\_proyecto FROM contrato INNER JOIN proyecto ON contrato.cod\_proyecto = proyecto.cod\_proyecto.

<sup>9</sup> SELECT contrato.\*, proyecto.nom\_proyecto, cliente.nom\_cliente, lote.valor\_lote, vendedor.nom\_vendedor FROM contrato, proyecto, cliente, lote, vendedor WHERE contrato.rut\_cliente = cliente.rut\_cliente AND contrato.cod\_proyecto = proyecto.cod\_proyecto AND lote.cod\_lote = contrato.cod\_lote AND lote.cod\_etapa = contrato.cod\_etapa AND contrato.cod\_vendedor = vendedor.cod\_vendedor AND contrato.en\_uso = 1.

demostrar si su *media* es igual a un cierto valor fijo [3,9]. Lo que permite aceptar o rechazar la hipótesis nula referente a la *media* de sus observaciones. El estadístico  $T = \min(T^+, T^-)$ , se construye a partir de las diferencias entre las observaciones y un valor fijo, que son ordenadas por rango<sup>10</sup> sin tomar en cuenta el signo<sup>11</sup>. En que  $T^+$  es la suma de los rangos asignados a las diferencias positivas y  $T^-$  es la suma de los rangos asignados a las diferencias negativas. La prueba consiste a rechazar la hipótesis nula  $H_0$ , a un nivel de significación  $\alpha$ , si: (1)  $T \leq T_{\alpha}$ , dada una hipótesis alternativa  $H_1: \mu_1 \neq \mu_0$ ; (2)  $T^- \leq T_{2\alpha}$  dada una hipótesis alternativa  $H_1: \mu_1 > \mu_0$ ; y (3)  $T^+ \leq T_{2\alpha}$  dada una hipótesis alternativa  $H_1: \mu_1 < \mu_0$ , donde  $T_{\alpha}$  es un valor crítico obtenido en tablas [3] para un valor  $n$ , que es el número de diferencias con rango asignado.

En la segunda parte de este artículo, se presenta la aplicación de la Prueba de T Wilcoxon a cada uno de los esquemas de consulta estudiados y definidos en el punto 2. En cuanto a la hipótesis alternativa está fue planteada de la forma (3) (ver más arriba) y se escogió un nivel de significación  $\alpha=0.05$ . Lo anterior implica que se rechaza la hipótesis nula al nivel 5% si  $T^+ \leq T_{2\alpha}$ , donde el valor crítico obtenido de tablas [3] para  $T_{0.1}$  es igual a 11, dado un tamaño de la muestra  $n = 10$  observaciones.

#### 4. Prueba T Wilcoxon aplicada a los esquemas anidados de consulta

Esto consiste en aplicar la Prueba de T Wilcoxon a cada uno de los esquemas de consulta del punto a) de la sección 2. Para ello, la tabla 1 muestra el tiempo de respuesta (en segundos) de la ejecución del esquema de consulta Not Exist (SELECT a.cod\_lote FROM lote a WHERE NOT EXISTS (SELECT b.cod\_lote FROM contrato b WHERE b.cod\_etapa = 31 AND b.cod\_proyecto = 14 AND a.cod\_lote = b.cod\_lote) AND a.cod\_etapa = 31 AND a.cod\_proyecto = 14) iterado en PostgreSQL.<sup>12</sup> En la misma tabla se indican las diferencias de tiempos de respuesta y los rangos asociados. Es importante destacar, tal como se explicó anteriormente, que las iteraciones en el cliente PostgreSQL se consiguen a partir de un ciclo while(rs.next()) en lenguaje Java. Los detalles del programa se pueden consultar en [4].

Anidado Servidor	Anidado Cliente	Diferencia (segs)	Rango
0.324	0.118	0.206	10
0.055	0.118	-0.063	7
0.059	0.118	-0.059	1.5
0.055	0.118	-0.063	7
0.058	0.119	-0.061	3.5
0.059	0.118	-0.059	1.5
0.055	0.118	-0.063	7
0.057	0.118	-0.061	3.5
0.056	0.118	-0.062	5
0.055	0.119	-0.064	9

Tabla 1. Tiempos de respuesta para un esquema anidado de consulta bajo PostgreSQL. Caso consulta Not Exist

En este caso, para poder realizar la prueba T Wilcoxon, se supuso [4] que las observaciones (tiempos de respuesta) tienen una distribución continua de igual media y que los pares de diferencias de tiempos de respuesta son independientes. Esto permite plantear las hipótesis nula ( $H_0$ ) y alternativa ( $H_1$ ) de la forma siguiente:

<sup>10</sup> Un valor racional positivo.

<sup>11</sup> Si la diferencia es cero entonces no hay rango asociado. En caso de haber varias diferencias iguales, a todas ellas se la asigna la media de los rangos que ocupen en conjunto.

<sup>12</sup> En ambos casos (Servidor/Cliente) se obtuvieron 18 tuplas retornadas.

- Ho: *Esquema anidado de consulta en el Servidor* posee igual media que *esquema anidado de consulta en el Cliente* bajo PostgreSQL.
- H<sub>1</sub>: *Esquema anidado de consulta en el Servidor* posee menor media que *esquema anidado de consulta en el Cliente* bajo PostgreSQL.

En la tabla 1 se puede observar que la suma de los rangos asignados a las diferencias positivas ( $T^+$ )<sup>13</sup> es igual a diez, por lo que se rechaza la hipótesis nula y pueden inferirse diferencias significativas en tiempos de respuesta para ese esquema de consulta iterado en el servidor bajo PostgreSQL. De igual forma se aplicó la prueba T Wilcoxon para el esquema anidado de consulta Not In. La tabla 2 resume el resultado de la prueba T Wilcoxon para ambas consultas.

Tipo de esquema	Tuplas retornadas	$T^+$	$T_{2\alpha}$	Rechazar Ho	<i>n</i>
Consulta Not Exist	18	10	11	SI	10
Consulta Not In	18	0	11	SI	10

Tabla 2. Prueba T Wilcoxon aplicada a los esquemas anidados de consulta

El resumen de la tabla 2, afirma que existe evidencia para rechazar la hipótesis nula y aceptar la hipótesis alternativa de menores tiempos de respuesta, al nivel significación del 5%, para un esquema anidado de consulta en el servidor sobre un esquema anidado de consulta en el cliente bajo PostgreSQL. Esto es válido para el tipo de consulta Not Exist y Not In.

## 5. Prueba T Wilcoxon aplicada a los esquemas iterados de consulta

Esto consiste en aplicar la Prueba de T Wilcoxon a cada uno de los esquemas de consulta del punto b) de la sección 2. Para ello, la tabla 3 muestra el tiempo de respuesta (en segundos) de la ejecución del esquema de consulta simple (SELECT DISTINCT rut\_cliente FROM contrato), tanto para MySQL como para PostgreSQL.<sup>14</sup> En la misma tabla se indican las diferencias de tiempos de respuesta y los rangos asociados.

MySQL	PostgreSQL	Diferencia (segs)	Rango
0.07	0.111	-0.041	10
0.07	0.11	-0.04	7
0.09	0.13	-0.04	7
0.07	0.11	-0.04	7
0.071	0.1	-0.029	2
0.07	0.101	-0.031	4
0.071	0.111	-0.04	7
0.07	0.11	-0.04	7
0.07	0.1	-0.03	3
0.08	0.1	-0.02	1

Tabla 3. Tiempos de respuesta para un esquema iterado de consulta. Caso consulta simple

Ahora tal como en los esquemas anidados, se supuso [4] que las observaciones (tiempos de respuesta) tienen una distribución continua de igual media y que los pares de diferencias de tiempos de respuesta son independientes. Esto permite plantear las hipótesis nula (Ho) y alternativa (H<sub>1</sub>) de la forma siguiente:

<sup>13</sup> Sin embargo,  $T^+=1.5*2+3.5*2+5+7*3+9=45$

<sup>14</sup> En ambos casos se obtuvieron 3583 tuplas retornadas.

Ho: *MySQL* posee igual media que *PostgreSQL* para cada tiempo de respuesta.

H<sub>1</sub>: *MySQL* posee menor media que *PostgreSQL* para cada tiempo de respuesta.

En la tabla 3 se puede observar que la suma de los rangos asignados a las diferencias positivas ( $T^+$ )<sup>15</sup> es igual a cero, por lo que se rechaza la hipótesis nula y pueden inferirse diferencias significativas en tiempos de respuesta para ese esquema de consulta simple. De igual forma se aplicó la prueba T Wilcoxon para los otros esquemas iterados de consulta (esquema de consulta condicional, esquema de consulta Inner Join y esquema de consulta condicional compleja). La tabla 4 resume el resultado de la prueba T Wilcoxon según el tipo de esquema iterado de consulta.

Tipo de esquema	Tuplas retornadas	$T^+$	$T_{2\alpha}$	Rechazar Ho	$n$
Consulta simple	3583	0	11	SI	10
Consulta simple condicional	3595	9	11	SI	10
Consulta Inner Join	3595	0	11	SI	10
Consulta condicional compleja	8408	0	11	SI	10

Tabla 4. Prueba T Wilcoxon aplicada a los esquemas iterados de consulta

El resumen de la tabla 2, la prueba T Wilcoxon afirma que existe evidencia para rechazar la hipótesis nula y aceptar la hipótesis alternativa de menores tiempos de respuesta, al nivel de significación del 5%, para MySQL.

## 6. Conclusiones

En este artículo se presenta un análisis estadístico no paramétrico, dado por la Prueba T Wilcoxon, que permite comparar, por una parte, los tiempos de respuesta de los esquemas anidados de consulta Not Exist y Not In bajo PostgreSQL, y por otra parte, los tiempos de respuesta de los esquemas iterados de consulta (simple, condicional, Inner Join y condicional compleja) ejecutados en MySQL y PostgreSQL respectivamente. A pesar de que MySQL superó en un 100% a PostgreSQL en todas las pruebas T Wilcoxon, éste no soporta la anidación de consultas SQL. Tal vez esto explica el hecho de que muchas aplicaciones Web utilizan a MySQL como repositorio de datos, ya que ellas requieren menores tiempos de respuestas. En cambio, PostgreSQL ha encontrado una gran aceptación en aplicaciones Cliente/Servidor orientadas a la gestión de datos que requieren menores tiempos de respuesta en la anidación de consultas SQL en el servidor. Sin embargo, tal como se advirtió al inicio del artículo, nada de esto es posible generalizarlo, puesto que se tomó en cuenta un sólo modelo de acceso remoto a datos perteneciente a una empresa en particular.

## 7. Trabajos futuros

Ampliar el análisis a otros modelos de datos, trasladando la lógica de mediciones propuesta en [4]. Así como también comparando otras funcionalidades de los SGBD, generando así nuevas variables de interés. Por ejemplo, el *uso de memoria*. El cual se define como la cantidad (en bytes) de memoria máxima o mínima utilizada para la ejecución de un esquema de consulta. En [4] se pueden consultar los resultados obtenidos al aplicar la Prueba T Wilcoxon para variaciones de memoria máxima y para variaciones de memoria mínima. Sin embargo, dado un

---

<sup>15</sup> Sin embargo,  $T=1+2+3+4+7*5+10=55$

recuestionamiento en la formulación de la hipótesis nula, no hemos estimado conveniente incluirlo en este artículo.

## **8. Agradecimientos**

Los autores expresan su gratitud al ingeniero Fernando San Martín Woerner por haber formado parte del equipo de investigación. De igual forma, se agradece al profesor Luis Cofré, de la Universidad Católica del Maule, por su colaboración en el análisis estadístico no paramétrico.

## **9. Referencias**

- [1] Bustos F., “Estudio Comparativo del Comportamiento de un Modelo de Datos en Tercera, Cuarta y Quinta Forma Normal”, Memoria, Universidad Católica del Maule, 2000.
- [2] Gardarin G., “Bases de données”, Eyrolles, 1999.
- [3] Gil J., “Estadística no Paramétrica”, RA-MA, Madrid, 1988.
- [4] Ibañez J., San Martín F., “Análisis de Consultas SQL bajo un ambiente de Programación JAVA”, Memoria, Universidad Católica del Maule, 2002.
- [5] Mallordy L., “Répartition d'objets dans les bases de données”, Hermès, Paris, 1995.
- [6] Urrutia A., San Martín F., Villarroel R., “Propuesta para el Tratamiento de Bases Datos Concurrentes usando una Herramienta Visual”, VII Encuentro Chileno de Computación 1999, Universidad Católica del Maule, 8 al 12 de Noviembre de 1999, Talca, Chile.
- [7] Urrutia A., Castro G., “Comparación de Metodologías de Diseño para generar Esquemas Relacionales”, VII Encuentro Chileno de Computación 1999, Universidad Católica del Maule, 8 al 12 de Noviembre de 1999, Talca, Chile.
- [8] Urrutia A., Jiménez L., “Una Propuesta para la Mantención de un Modelo de Datos en 3FN, utilizando FNDC”, VI Encuentro Chileno de Computación 1998, Universidad Católica del Norte, 12 al 14 de Noviembre de 1998, Antofagasta, Chile.
- [9] Wilcoxon F., Wilcoxon R., “Some Rapid approximate Statistical Procedures”, American Cyanamid Company, Pearl River, New York, 1964.