Comportamiento de Bases de Datos No Relacionales en Entornos Distribuidos

Ana Lía Carabio[1], Marcelo G. Benedetto[1], Marcelo A. Falappa[2]

[1]Facultad de Ciencias de la Administración - Universidad Nacional de Entre Ríos
Monseñor Tavella 1424 – Concordia, Entre Ríos (3200) - Tel.: +54(0345)4231406
{anacar, marben}@fcad.uner.edu.ar

[2]Departamento de Ciencias e Ingeniería de la Computación - Universidad Nacional del Sur
Avenida Alem 1253 - Bahía Blanca (B8000CPB) - Tel.: +54(0291)4595135
mfalappa@cs.uns.edu.ar

Resumen

La estructura de un sistema de información típico actual consta, en general, de programas de aplicación distribuidos, un Sistema de Gestión de Base de Datos (SGBD) y una red que permite entregar la información desde y hacia los distintos usuarios.

Dentro de los SGBD actuales han surgido las bases de datos NOSQL (Not Only SQL) y los Sistemas de Almacenamiento de Datos Masivos (*Big Data Storage Systems*), que almacenan la información en forma distribuida y permiten una mayor escalabilidad.

El tiempo utilizado en la comunicación entre los distintos nodos afecta la performance de la base de datos y, por consiguiente, el tiempo de respuesta al usuario, poniendo en evidencia la importancia del comportamiento de la red en el rendimiento de un sistema de información.

En este trabajo de investigación se buscará analizar el comportamiento de una base de datos no puramente relacional, y el comportamiento de la red de comunicaciones, con el fin de evaluar el rendimiento a medida que se actualiza la información en un entorno distribuido.

Palabras clave: Bases de Datos, Sistemas Distribuidos, Redes de Comunicación, Medidas de Performance.

Contexto

Este trabajo se desarrolla dentro del Proyecto de Investigación y Desarrollo PID 7042 "Estudio Comparativo y Análisis de Rendimiento de los Lenguajes de Manipulación de Datos en Bases de Datos Orientadas a Objetos y Bases de Datos Objeto-Relacionales"[1], cuyo período de ejecución será desde noviembre de 2014 a noviembre de 2017, en el marco de un Acuerdo de Colaboración Académico-Científico entre la Facultad de Ciencias de la Administración de la Universidad Nacional de Entre Ríos (UNER) y el Instituto de Ciencias e Ingeniería de la Computación (ICIC) del Departamento de Ciencias e Ingeniería de la Computación

(DCIC) de la Universidad Nacional del Sur (UNS).

Uno de los objetivos del proyecto apunta a establecer comparaciones en el rendimiento de sistemas desarrollados en lenguajes orientados a objetos que interactúan con diversos modelos de bases de datos.

Además, este proyecto prioriza la formación de recursos humanos para investigación en la Facultad de Ciencias de la Administración de la UNER, especializados en la línea de investigación denominada "Ingeniería de Software y Lenguajes de Programación" establecida por Res. 25/11 del Consejo Directivo.

Introducción

Las redes de computadoras y las bases de datos son utilizadas masivamente por las aplicaciones de hoy en día, dado que la disponibilidad y acceso a la información se ha convertido en una herramienta indispensable para la toma de decisiones.

La estructura de un sistema de información típico actual consta, en general, de programas de aplicación distribuidos, un SGBD y una red que permite entregar la información desde y hacia los distintos usuarios.

Dentro de los SGBD, el modelo relacional ha sido el más utilizado, a pesar de ciertas limitaciones. En la actualidad, existen extensiones de estos sistemas que incorporan los conceptos de tipos complejos y orientación a objetos conformando los SGBDOR *Sistemas de Gestión de Base de Datos Objeto-Relacionales* (SGBDOR) y los S*istemas de Gestión de Base de Datos Orientados a Objetos* (SGBDOO) [2]. También han surgido las bases de datos NOSQL (*Not Only SQL*) [2], las cuales permiten mayor escalabilidad que los sistemas tradicionales [3], y los Sistemas de Almacenamiento de Datos Masivos (*Big Data Storage Systems*). Estos sistemas han emergido en compañías y organizaciones que almacenan grandes cantidades de información en forma distribuida, entre las cuales podemos citar Google, Yahoo, Amazon, Facebook, etc.

En cuanto a la tecnología de redes de computadoras, la misma promueve una forma de trabajo que procura, en general, evitar la centralización; mientras que uno de los objetivos del uso de las bases de datos es la necesidad de integrar los datos y proveer mecanismos que controlen el correcto acceso a los mismos [4].

En los sistemas distribuidos los datos se encuentran almacenados en varios servidores, y los clientes pueden acceder a ellos por medio de una red de comunicación [5, 6]. Cuando el SGBD y/o los datos se encuentran alojados en sitios remotos, las consultas a la base de datos pueden generar que la información atraviese varias redes y/o dispositivos (nodos) hasta llegar a destino, lo que incide directamente en el rendimiento.

Independientemente de la red utilizada, la transferencia de información a través de cualquier canal de comunicación requiere de un tiempo que dependerá del ancho de banda del canal, de la longitud y sobrecarga del enlace, de la velocidad y eficiencia de la red, del número de nodos, entre otros [6, 7]. El tiempo utilizado en la comunicación constituye un retardo que afecta la performance de la base de

datos y, por consiguiente, el tiempo de respuesta al usuario [8].

Las bases de datos relacionales y objeto-relacionales fueron concebidas, inicialmente, para instalaciones centralizadas, aunque conceptualmente pueden implementarse de manera distribuida [8]. Cuando las aplicaciones se ejecutan en un entorno distribuido, tanto el programador de aplicación como el usuario del sistema deberían independizarse de aspectos físicos tales como: cantidad de nodos de la red, topología de la misma, réplicas de los datos, y/o fragmentación de las relaciones. No obstante, si nos abstraemos totalmente de la red y no consideramos que los recursos están físicamente ubicados en varias computadoras [4], se corre el riesgo de caer en los "famosos" supuestos de la computación distribuida, que terminan siendo falsos y generan grandes problemas a largo plazo: la red es confiable, la latencia es cero, el ancho de banda es infinito, la red es segura, la topología no cambia, existe un solo administrador, el costo de transporte es cero y la red es homogénea [9].

El rendimiento de un SGBD distribuido se ve afectado por el número de sitios en los que dicho sistema está distribuido, así como también por el grado de replicación de los datos, entre otros parámetros [8,10,12]. Por ejemplo, en una base de datos relacional tradicional, las relaciones se pueden fragmentar horizontalmente (por tuplas), verticalmente (a través de la descomposición en subesquemas), y combinado.

En particular, las bases de datos NOSQL permiten una replicación entre nodos más simple que la de las bases de datos relacionales, utilizando principalmente dos técnicas (replicación y sharding), de las que derivan otros modelos de distribución. La replicación duplica los datos en múltiples nodos, utilizando el modelo maestro-esclavo o el modelo peer-to-peer, mientras que sharding divide (fragmenta) la información en varios nodos. Estas técnicas, con sus ventajas y desventajas, también pueden combinarse entre sí, generando un esquema más complejo [11].

Líneas de Investigación, Desarrollo e Innovación

En la actualidad han cobrado importancia las bases de datos no puramente relacionales, caracterizadas, principalmente, por su almacenamiento distribuido y su fácil escalabilidad. En esta línea, se buscará analizar el comportamiento de bases de datos no relacionales del tipo NOSQL con distintos esquemas de distribución, con la finalidad de evaluar su rendimiento a medida que se actualiza la distribución de la información.

A su vez, la distribución de información calificada, como pueden ser los datos multimediales, se encuentra con ciertos obstáculos que presentan las redes de comunicaciones para soportar este tipo de datos. En particular, la transmisión de ciertos datos requiere un ancho de banda determinado y no deberían perder continuidad. Por ejemplo, cuando se transfiere audio y/o video mediante *streaming*, es necesario que la información tenga continuidad de

reproducción. En este sentido, se intentará formular y/o adaptar protocolos de comunicación que permitan garantizar que la información transferida a través de la red tenga la fluidez esperada.

Resultados y Objetivos

Dada la incidencia que indudablemente tiene el comportamiento de la red sobre el rendimiento global de un sistema de información y de una base de datos distribuida no puramente relacional en particular, se hace necesario evaluar el desempeño de la red y posteriormente determinar la incidencia de dicho comportamiento sobre el rendimiento de la base de datos. Para ello se prevé:

- Seleccionar e instalar distintas bases de datos del tipo NOSQL, con distintos modelos de distribución, sobre una red con al menos dos o tres computadoras personales con procesadores de varios núcleos, y un número incremental de máquinas virtuales.
- Evaluar el desempeño de la red en distintas instancias de trabajo, utilizando como indicadores, inicialmente, la tasa de transferencia (*throughput*), el retardo y la pérdida de paquetes. En caso que la investigación así lo requiera, se estimará la incorporación de otras variables de medición, y/o determinar preferencias entre las variables ya medidas.
- Evaluar el comportamiento de las distintas bases de datos NOSQL instaladas y determinar la incidencia del comportamiento de la red sobre el rendimiento de dichas bases de datos.

Finalmente, se estudiarán los protocolos distribuidos existentes y, si fuera necesario, se reformularán y/o generarán nuevos protocolos con el fin de lograr un mejor comportamiento.

Los resultados de esta investigación dependerán de: las variables de medición contempladas, los datos manipulados por las bases de datos no relacionales, la cantidad de lecturas y escrituras realizadas en la misma, y la cantidad de mensajes de control transferidos a través de la red.

Formación de Recursos Humanos

Como parte del actual proyecto de investigación se espera que uno de los docentes investigadores, y que es autor de este artículo, complete su Tesis de Magister en Redes de Datos en la Facultad de Informática de la Universidad Nacional de La Plata. A su vez, se buscará formar nuevas sublíneas de investigación relacionadas a este proyecto, así como también la formación de nuevos alumnos en los posgrados dictados en el ámbito de la UNER y de la UNS.

Referencias

[1] Benedetto, Marcelo G., Carabio, Ana Lía R., Alvez, Carlos E., Fernández, Miguel, Etchart, Graciela, Cabrera, Sergio A., Benítez, Horacio D., Falappa, Marcelo A, Martínez, Diego C. & Cobo, M. Laura (2015). *Selección de lenguajes orientados a objetos para un estudio comparativo y análisis de rendimiento*. En **XVII Workshop de Investigadores**

en Ciencias de la Computación, WICC'2015.

[2] Elmasri, Ramez, & Navathe, Shamkant B. (2015). *Fundamentals of Database Systems.* 7th. Edition. **Addison Wesley**.

[3] Cattell, Rick (2011). *Scalable SQL and NoSQL data stores*. **ACM SIGMOD Record**, 39(4), 12-27.

[4] Özsu, M. Tamer, & Valduriez, Patrick (2011). *Principles of Distributed Database Systems.* Third Edition. **Springer Science & Business Media**.

[5] Kurose, James F. & Ross, Keith W. (2012). *Computer Networking: A Top-Down Approach.* Sixth Edition. **Pearson**.

[6] Sahu, Amir. K., & Hemrajani, Naveen (2012). *An Analysis of Distributed Computer Network Administration*. **International Journal of Computer Technology and Applications**, *3*(2): 660-667.

[7] Gámiz Caro, Juan, & Martínez García, Herminio (2008). *El retardo del mensaje en sistemas de control distribuidos a través de Ethernet estándar.* In la Quinta **Conferencia Internacional de la Facultad de Ingeniería Eléctrica**: 1-7.

[8] Silbertschatz, Abraham. & Korth, Henry. Sixth Edition (2010). *Database System Concepts*. **McGraw-Hill Education**.

[9] Rotem-Gal-Oz, Arnon. (2006). *Fallacies of distributed computing explained*. *http://www.rgoarchitects.com/ Files/fallacies. pdf*.

[10] Maabreh, Khaled S. (2011). An *Analyzing Study of the Distributed Database System Parameters*. Technical Report. Faculty of Science and Information technology. Zarqa University. Jordan, Al Zarqa.

[11] Pérez Blanco, Carlos. (2013). *NoSQL databases in cross-platform development*.

[12] Garcia-Molina, Hector, Ullman, Jeffrey D. & Widom, Jennifer (2008). *Database Systems: The Complete Book.* Second Edition. **Pearson**.