NoSQL: MODELOS DE DATOS Y SISTEMAS DE GESTIÓN DE BASES DE DATOS

Silvina Migani[1], Cristina Vera[1], María Inés Lund[2]

[1]Departamento de Informática - [2]Instituto de Informática, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan
silvina.migani@gmail.com; civerados@gmail.com; mlund@iinfo.unsj.edu.ar

RESUMEN

Los Sistemas de Gestión de Bases de Datos (SGBDs) NoSQL surgieron como una alternativa de solución a problemas no resueltos eficientemente por los SGBDs tradicionales. Sin embargo, a diferencia de las bases de datos relacionales, el término NoSQL comprende múltiples modelos de datos, cada uno con características diferentes. Además, hoy en día existen muchos SGBDs para cada modelo de esta nueva familia de bases de datos. Por tal motivo, se considera valioso profundizar en el estudio de este nuevo paradigma, haciendo hincapié en la comparación cualitativa de los diferentes modelos y también cuantitativa de distintos gestores de bases de datos NoSQL a través de un benchmark que establezca puntos de referencia. Así, los datos generados por este estudio permitirán orientar a los usuarios en la elección del SGBD apropiado ante un problema específico.

CONTEXTO

Este trabajo forma parte del proyecto "NoSQL: UN NUEVO PARADIGMA EN BASES DE DATOS", de carácter bi-anual, presentado en convocatoria de la UNSJ en Diciembre de 2017. Se encuentra, al momento de esta presentación, en evaluación por pares externos y se espera su aprobación.

El proyecto se encuentra estrechamente relacionado con las dos asignaturas de Bases de Datos de las carreras de Licenciatura en Ciencias de la Computación y Licenciatura en Sistemas de Información de la Universidad Nacional de San Juan (Base de Datos y Tópicos Avanzados de Base de Datos) y con la asignatura Ingeniería de Software II (también de ambas carreras). El equipo de trabajo está formado por docentes investigadores del Departamento y del Instituto de Informática. Esta conjugación fortalecerá fundamentalmente el análisis y la comparación de los diferentes modelos existentes dentro de este nuevo paradigma.

1. INTRODUCCIÓN

Base de Datos NoSQL

Carlo Strozzi en 1998, utilizó por primera vez la expresión NoSQL para referirse a una base de datos open-source relacional, que prescindía del lenguaje SQL [1]. Claramente no es el significado que en la actualidad se le da a dicho término. Luego, Eric Evans, en 2009, reintrodujo el término para referirse a bases de datos no relacionales, distribuidas, linealmente escalables, de código abierto y que no garantizaban las tradicionales propiedades ACID (Atomicity – Consistency – Isolation - Durability) [2].

Actualmente, si bien se sigue discutiendo sobre la conveniencia o no del nombre, es globalmente conocido y aceptado que el término NoSQL refiere a bases de datos de esquema flexible, distribuidas y que no se ajustan al modelo tradicional de transacciones ACID [3].

En el ámbito de un simposio de "Principios de Computación Distribuida" de ACM en el año 2000, Eric Brewer enuncia el teorema de CAP, donde manifiesta que todo sistema distribuido con datos compartidos no puede satisfacer simultáneamente las tres características siguientes: Consistencia, Alta Disponibilidad y Tolerancia a Particiones [4,5,6]. Consecuentemente, los SGBDs NoSQL al no poder satisfacer en su totalidad las tres propiedades, adhieren al modelo BASE (Basically Available, Soft State, Eventually Consistent) [7]. Es decir, básicamente disponible, de estado flexible y eventualmente consistente.

Según [3] los sistemas NoSQL mantienen seis propiedades fundamentales:

1. Habilidad de escalar horizontalmente sobre muchos servidores.
2. Habilidad de replicar y distribuir datos (particiones) en muchos servidores.
3. Interfaz o protocolo simple a nivel de llamada (en contraste con el enlace de SQL).
4. Modelo de concurrencia más débil que el ACID (típico de los sistemas de bases de datos relacionales).
5. Uso eficiente de RAM e índices distribuidos.
6. Habilidad de agregar dinámicamente nuevos atributos a los registros de datos.

Varios autores [3][8] clasifican los sistemas de bases de datos NoSQL según el modelo de datos subyacente en:

- Clave-Valor: Sistemas que almacenan valores vinculados a un índice que se basa en una clave definida por el programador.
- De Documentos: Sistemas que almacenan documentos y proveen un lenguaje de consulta. Además, los documentos pueden ser indexados.
- De Registros Extensibles u Orientados a Columnas: Estos sistemas almacenan registros extensibles que pueden particionarse vertical u horizontalmente en diferentes nodos de una red.

Benchmarks

Desde los inicios de la informática, los benchmarks han sido ampliamente utilizados como marco de referencia para obtener resultados objetivos al comparar sistemas. En general, efectuar las pruebas no es una tarea sencilla, requiere consideraciones meticulosas e iteraciones para poder obtener resultados certeros. Actualmente, existe una amplia gama de benchmarks para evaluar el rendimiento de diferentes componentes de software y hardware y dentro del dominio de los sistemas NoSQL se han propuesto varios. La elección de un benchmark específico debiera considerar los siguientes criterios fundamentales: relevancia dentro de su dominio, simplicidad y escalabilidad [10].

La aparición de Internet generó grandes cambios en la informática. Las aplicaciones comenzaron a evolucionar y a demandar nuevos requerimientos que los SGBDs tradicionales no pudieron satisfacer adecuadamente. En consecuencia, surgieron nuevos modelos de bases de datos y gestores que los soportaron: bases de datos de documentos, clave-valor, orientados a columnas; entre los más destacados [3, 8], enunciados anteriormente. Es así que

actualmente, bajo el paraguas NoSQL, coexiste una gran variedad de sistemas de bases de datos. Por eso, se considera valioso poder identificar, distinguir y caracterizar estos diferentes enfoques; además de evaluar la performance de diferentes gestores de bases de datos. El conocimiento servirá de guía en futuros problemas a resolver en escenarios concretos.

Las actividades a desarrollar son las siguientes:

- Estudio exploratorio: Se definirá la estrategia de búsqueda y luego, ejecutará el mapeo sistemático de la literatura [8] que permita recolectar información sobre sistemas de bases de datos NoSQL y sobre los benchmarks existentes dentro de esta temática.
- Estudio y ensayo de SGBDs NoSQL: Luego de identificar algunos sistemas de gestión destacados, se procederá a estudiarlos y realizar ensayos que permitan experimentar diferencias y semejanzas.
- Selección, estudio y utilización de un benchmark apropiado para la comparación cuantitativa (tiempo de respuesta) de diferentes SGBDs NoSQL.
- Análisis de los datos recogidos, comparando cualitativa y cuantitativamente los diferentes sistemas de gestión de bases de datos NoSQL.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La línea de investigación seguida por el proyecto marco de este trabajo, es el estudio y comparación de modelos y gestores de bases de datos NoSQL.

3. RESULTADOS Y OBJETIVOS

Se pretende alcanzar un amplio conocimiento de conceptos relativos al paradigma NoSQL:

- Necesidades de su utilización
- Características relevantes y distintivas
- Diferenciación con las bases de datos relacionales
- Taxonomías existentes
- Caracterización de cada tipo identificado

Y, posteriormente, con el conocimiento adquirido:

- Hacer un estudio exploratorio de los SGBD NoSQL existentes.
- Identificar sistemas de gestión destacados y realizar ensayos.
- Analizar diferentes benchmarks.
- Elegir un benchmark que permita realizar una comparación cuantitativa de los sistemas de bases de datos elegidos.
- Hacer un estudio comparativo cualitativo y cuantitativo (tiempo de respuesta) de diferentes sistemas de gestión de bases de datos NoSQL.

4. FORMACIÓN DE RECURSOS HUMANOS

Este proyecto contribuirá a la profundización y consolidación del conocimiento de esta área temática por parte de cada uno de sus integrantes. Además, se pretende generar distintos temas para trabajos finales de grado de las carreras del Departamento y también de postgrado. De hecho, dos de las integrantes de este proyecto, durante la vigencia del mismo, realizarán su tesis de Maestría en esta temática. Además, es de interés del grupo proponer alumnos para becas de investigación.

5. BIBLIOGRAFÍA

[1] Strozzi, C. (1998). NoSQL-A relational database management system. Lainattu, 5, 2014.

[2] Evans, E. (2009). NoSQL: What's in a name. Eric Evans's Weblog. http://blog. sym-link. com/2009/10/30/nosql_whats_in_a_ name. html. [Ultimo Acceso: Diciembre de 2017].

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

[4] Brewer, E., & Fox, A. (1999). Harvest, yield, and scalable tolerant systems. In HOTOS Proceedings of the The Seventh Workshop on Hot Topics in Operating Systems.

[5] Brewer, E. (2010). A certain freedom: Thoughts on the CAP theorem. 29th ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing PODC, 335–335.

[6] Abadi, D. (2012). Consistency tradeoffs in modern distributed database system design: CAP is only part of the story. Computer, 45(2), 37-42.

[7] Pritchett, D. (2008). Base: An acid alternative. Queue, 6(3), 48-55.

[8] Makris, A., Tserpes, K., Andronikou, V., & Anagnostopoulos, D. (2016). A classification of NoSQL data stores based on key design characteristics. Procedia Computer Science, 97, 94-103.

[9] Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology, 64, 1-18.

[10] Gray, J. (1992). Benchmark handbook: for database and transaction processing systems. Morgan Kaufmann Publishers Inc..