Benchmarking de Bases de Datos NoSQL para el almacenamiento de Modelos Semánticos

Nélida Raquel Cáceres, Ana Carolina Tolaba, Ricardo Daniel Pérez, Jairo Joel Maximiliano Quispe, Cintia Silvana Rodríguez, Iván Leandro Sandoval

Facultad de Ingeniería - Universidad Nacional de Jujuy
Ítalo Palanca 20 San Salvador de Jujuy – 0388 4221576
nrcaceres@fi.unju.edu.ar

RESUMEN

Las bases de datos tradicionales presentan limitaciones al momento de almacenar y procesar los datos debido a la cantidad (datos no sólo generados de transacciones sino desde sensores, dispositivos móviles o clics de la web) y estructura de los mismos (por ejemplo datos de redes sociales, datos espaciales o datos semánticos). La solución a éste último concepto es el surgimiento de una nueva tecnología denominada sistema de gestión de bases de datos no relacionales (NoSQL).

Los sistemas NoSQL se emplean cada vez más para el manejo de datos semánticos, es decir, modelos de datos que incluyen información semántica. Sin embargo, todavía es difícil comprender sus principales ventajas y desventajas en este contexto.

Este proyecto tiene la intención de caracterizar y comparar mediante un proceso de benchmarking bases de datos NoSQL orientadas a grafos para el almacenamiento y procesamiento de datos semánticos. El objetivo de la evaluación no es definir cuál es mejor, sino determinar aspectos comunes, características de consulta, e identificar las diferencias entre los sistemas NoSQL.

Palabras clave: Bases de Datos NoSQL, Modelo Semántico, Benchmarking.

CONTEXTO

Este trabajo es financiado por la Facultad de Ingeniería de la Universidad Nacional de Jujuy, en el contexto de la convocatoria realizada para Proyectos orientados a la Investigación Básica y a la Investigación Aplicada en el Área de Informática de la Facultad de Ingeniería. La convocatoria fue realizada con el fin de fortalecer la formación de recursos humanos en investigación. El proyecto de investigación tiene previsto una duración de un año, desde enero hasta diciembre del 2018, con la posibilidad de solicitar su extensión de acuerdo a los resultados obtenidos.

1. INTRODUCCIÓN

En la actualidad los volúmenes de datos que se generan y consumen se destacan por un crecimiento acelerado, lo que implica que los repositorios que contienen una colección de datos, no sólo sean considerados para realizar las consultas tradicionales, sino también como repositorios a partir de los cuales se puede obtener información relevante que sea útil para la toma de decisiones.

El modelo tradicional de base de datos relacional (RDBM - Relational Database Management) es consistente, y sus ventajas y desventajas son bien conocidas [1]. Sin embargo, este modelo presenta limitaciones cuando la interconectividad de los datos es importante, debido a que la manipulación de los datos en una base de datos relacional puede ser más compleja y consumir más tiempo. Esto ha llevado a emplear nuevos enfoques bajo el concepto de sistemas de bases de datos NoSQL (Not Only SQL), los cuales permiten gestionar el volumen de datos en constante aumento. Las bases de datos NoSQL son consideradas de próxima generación y se caracterizan por ser no relacionales, distribuidos, de código abierto y escalables horizontalmente [2, 3].

NoSQL corresponde a una estrategia de persistencia que no sigue el modelo de datos relacional, y que no utiliza SQL como lenguaje de consulta [4], en otras palabras, no están supeditadas a una estructura de datos en forma de tablas y relaciones entre ellas, permitiendo a los usuarios almacenar información en formatos diferentes a los tradicionales. Algunas aplicaciones de estas bases de datos pueden observarse en [5, 6, 7].

Las bases de datos NoSQL ofrecen diferentes modelos de datos como [8, 9]:

- Wide-Column Store, a diferencia de las bases de datos relacionales los nombres y el formato de las columnas puede variar de una fila a otra dentro de la misma tabla;
- Document Store, orientados a almacenar datos semiestructurados y permiten realizar consultas por atributos presentes en sus valores;
- Key-Value Stores, implementan una clave para la indexación y recuperación de datos, usualmente implementadas de manera distribuida;
- Graph, diseñadas para datos que consisten en entidades interconectadas con un número finito de relaciones entre ellos. Representan la información como un grafo usando vértices y aristas.

Las bases de datos NoSQL orientadas a grafos han cobrado mayor importancia en la actualidad debido al crecimiento en los proyectos que necesitan de una base de datos donde la importancia de la información depende de la relaciones, por ejemplo, web semántica [10], web mining [11] entre otros. Estas relaciones se implementan a través de modelos semánticos, los cuales permiten captar mejor el significado (semántica) de los datos contenidos en una base de datos. El objetivo de los modelos de datos semánticos es capturar el significado de los datos mediante la integración de conceptos relacionales con conceptos de abstracción más poderosos. El modelo semántico más empleado en los últimos años es el modelo ontológico.

Los sistemas NoSQL se emplean cada vez más para el manejo de datos semánticos. Sin embargo, todavía es difícil comprender sus principales ventajas y desventajas en este contexto.

Este proyecto tiene la intención de caracterizar y comparar mediante un proceso de benchmarking, bases de datos NoSQL orientadas a grafos para el almacenamiento y procesamiento de datos semánticos. El objetivo de la evaluación no es definir cuál es mejor, sino determinar aspectos comunes, características de consulta, e identificar las diferencias entre los sistemas NoSQL.

2. LÍNEAS DE INVESTIGACIÓN y DESARROLLO

En este proyecto de investigación además del estudio y profundización de los

conceptos inherentes a grafos, información semántica y su forma de almacenamiento, se propone comparar el desempeño de diferentes bases de datos NoSQL orientadas a grafos a través de un proceso de benchmarking. El objetivo de este proceso es generar conocimiento especializado en el área de modelado, almacenamiento y procesamiento de información semántica en lo referente a la representación y el uso del conocimiento en sistemas computacionales.

3. RESULTADOS Y OBJETIVOS

De acuerdo a las actividades planificadas para el desarrollo del proyecto, se espera:

- Recopilar información referente a bases de datos NoSQL orientada a grafos para determinar aspectos comunes, características de consulta y diferencias entre los sistemas NoSQL.
- Analizar la forma de empleo de las bases de datos NoSQL orientada a grafos.
- Elaborar y difundir los resultados de la experiencia a través de publicaciones y presentación en congresos y eventos.
- Fortalecer las capacidades de los recursos humanos (alumnos participantes) en actividades de investigación.
- Contribuir a la definición de futuros temas de tesinas de grado para los alumnos participantes del proyecto de investigación.

4. FORMACIÓN DE RECURSOS HUMANOS

El proyecto está siendo desarrollado por un equipo conformado por docentes investigadores del Grupo de Investigación y Desarrollo en Ingeniería de Software (GIDIS) de la Facultad de Ingeniería de la Universidad Nacional de Jujuy, a continuación se detallan los responsables del proyecto:

- La Ing. Nélida Raquel Cáceres (Directora, Categoría de Investigación IV) quien coordina las actividades del proyecto y dirige a los integrantes del equipo. Actualmente realizando tesis de maestría vinculada al área de bases de datos.
- La Ing. Ana Carolina Tolaba (Codirectora, Categoría de Investigación V) y dirige a los integrantes del equipo. Actualmente realizando tesis de doctorado vinculada al área de modelado conceptual de datos a través de modelos semánticos.

Participan del proyecto alumnos avanzados de la carrera de Ingeniería Informática:

- Ricardo Daniel Pérez
- Jairo Joel Maximiliano Quispe
- Cintia Silvana Rodriguez
- Iván Leandro Sandoval

Con la realización de este proyecto de investigación se espera la consolidación de los miembros del grupo en especial de los alumnos como jóvenes investigadores. Además, el proyecto brindará un marco propicio para la iniciación de trabajos finales de grado de la carrera Ingeniería Informática.

5. BIBLIOGRAFIA

[1] Miller, J. J. “Graph database applications and concepts with neo4j”. In Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA. 2013. Vol. 2324, p. 36.

[2] Sitalakshmi Venkatraman, K. F., Kaspi, S., & Venkatraman, R. "SQL Versus NoSQL Movement with Big Data Analytics". I.J. Information Technology and Computer Science, 2016, p. 59-66.
[3] "NOSQL Databases," http://nosql-database.org Acceso: Octubre, 2017.
[4] Arévalo, H. H. R., & Cubides, J. F. H. "Un viaje a través de bases de datos espaciales NoSQL". Redes de ingeniería, 2013, 4(2), pp. 57-69.
[5] Rodríguez Pérez, A., Rodríguez Hernández, D., & Díaz Martínez, E. "Selección de Base de Datos No SQL para almacenamiento de Históricos en Sistemas de Supervisión". Revista Cubana de Ciencias Informáticas, 10(3), 159-170, (2016).
[6] Martín, A., Chávez, S. B., Rodríguez, N. R., Valenzuela, A., & Murazzo, M. A. "Bases de datos NoSQL en cloud computing". In XV Workshop de Investigadores en Ciencias de la Computación. (2013, June).
[7] Valenzo, M. R., Valencia, R. E. C., & Castro, J. M. M. "Integración de búsquedas de texto completo en Bases de Datos noSQL". Revista Vínculos, 8(1), 80-92. (2013)
[8] Leavitt, N. "Will NoSQL Databases Live Up to Their Promise?". IEEE Computer, vol. 43, no. 2, pp. 12–14, 2010. Available on: http://www.leavcom.com/pdf/NoSQL.pdf
[9] Sadalage, P. J., & Fowler, M. "NoSQL distilled: a brief guide to the emerging world of polyglot persistence". Pearson Education. Upper Saddle River, NJ. 2012.
[10] Hayes and Gutierrez C., "Bipartite Graphs as Intermediate Model for RDF" in Proceedings of the 3th International Semantic Web Conference (ISWC). LNCS, no. 3298. Springer-Verlag, Nov 2004, pp. 47–61.
[11] Schenker A., Bunke H., Last M., and Kandel A., "Graph-Theoretic Techniques for Web Content Mining". Series in Machine Perception and Artificial Intelligence. World Scientific, 2005, vol. 62.