Informe de Proyecto de Fin de Curso

Aldo Díaz Betizagasti

***Resumen*—En este trabajo, se lleva a cabo un experimento que se trata de realizar pruebas de performance entre distintas opciones de bases de datos no relacionales, con el foco principalmente en las operaciones de lectura y escritura. Se trata de replicar el experimento realizado en: *Which NoSQLDatabase? A Performance Overview"* para obtener una referencia actualizada con las versiones más actuales de las herramientas.**

I. Introducción

Las bases de datos son un conjunto de información o datos organizados y estructurados que son almacenados virtualmente en un sistema informático. Cumplen un rol fundamental en prácticamente cualquier sistema de software al día de hoy, tanto a nivel personal como organizacional. Desde los inicios, las bases de datos han sido estructuradas en tablas y relaciones. Sin embargo, sus capacidades se ven limitadas con el constante avance de las necesidades en la capacidad de almacenamiento y el análisis de la información. En respuesta a ello, surgen nuevos modelos de bases de datos que rompen el paradigma tradicional. En este trabajo, se lleva a cabo un experimento que se trata de realizar pruebas de performance entre distintas opciones de bases de datos no relacionales, con el foco principalmente en las operaciones de lectura y escritura. A su vez, el mismo se realiza dentro del marco de la materia: Base de Datos No Relacionales que se dicta en la Facultad de Ingeniería en la Universidad de la República.

II. Trabajos Relacionados

Este trabajo está fuertemente basado en *"Which NoSQL Database? A Performance Overview"* de autores: Veronika Abramova, Jorge Bernardino y Pedro Furtado. Se trata de poder replicar el experimento realizado en el mismo utilizando los productos y herramientas de la actualidad y así poder sacar conclusiones en base a los resultados obtenidos.

III. Desarrollo

En el trabajo de referencia se tienen en cuenta 5 bases de datos al momento de realizar la comparación: Cassandra, HBase, MongoDB, OrientDB y Redis. Estas bases de datos no relacionales se clasifican dentro de tres categorías distintas: Familia de columnas (Cassandra y HBase), Documentos (MongoDB y OrientDB) y Clave-valor (Redis). Sin embargo, en este trabajo se incluyeron cuatro bases de datos de las nombradas: Cassandra, MongoDB, OrientDB y Redis, sin embargo se mantienen presente las 3 distintas categorías.

Para la ejecución de las pruebas de evaluación se utiliza la herramienta Yahoo Cloud Serving Benchmark (YCSB), que abstrae varios aspectos como la definición de algoritmos, medición de tiempos, carga de datos, entre otros. Para utilizar esta herramienta, se debe de configurar la base de datos en el ambiente en cuestión, configurar la integración con YCSB, definir cargas de trabajos y realizar la carga y ejecución de pruebas. También, es imprescindible configurar un conjunto de reglas que determinan las cargas de trabajo, éstas constan en definir cuáles son las operaciones a realizar, cuántas operaciones serán realizadas, definir las proporciones entre los tipos de operaciones a realizar, entre otros.

Las pruebas se ejecutaron en dos ambientes distintos, debido a que una de las bases de datos no tiene soporte para el ambiente configurado. Las pruebas para la base de datos de Cassandra, MongoDB y OrientDB fueron realizadas en una máquina con el sistema operativo Windows 10 Pro 64bits, 16GB RAM, procesador i5 4690 3.50Ghz y las pruebas sobre la base de datos Redis se realizaron sobre la misma máquina pero en un ambiente de Ubuntu utilizando la tecnología de WSL[1] para windows. (capaz que agregar que dado el nivel de la pc, los recursos sobran, pero capaz que tendría que agregar un link)

IV. Cargas de trabajo

Se definieron diversas cargas de trabajo con el objetivo de probar las operaciones de lectura y escritura en las distintas bases de datos mencionadas. Como se mencionó anteriormente, las cargas de trabajo son una parte principal que se debe de realizar en la herramienta de Yahoo Cloud Serving Benchmark para realizar las pruebas en la que se define el tipo de operaciones a ser ejecutados, las proporciones de las mismas si hay más de un tipo y la cantidad. Más allá de los mencionados, también existen otros aspectos parametrizables de las cargas de trabajo pero que no se enfocan directamente con el objetivo del trabajo, por lo que se los dejan constantes y con sus valores definidos por defecto.

En particular, se definieron tres tipos de cargas de datos:

- Carga de trabajo uno: 50 % de la operaciones son de lectura y 50 % de las operaciones son de escritura.
- Carga de trabajo dos: 100 % de las operaciones son de lectura.
- Carga de trabajo tres: 100 % de las operaciones son de escritura.

En total, todas las cargas de trabajo cargan 600.000 registros en las distintas bases de datos y se realizan sobre ese conjunto un total de 1000 operaciones.

V. Ejecución de pruebas

En la figura uno, se puede apreciar los resultados obtenidos tras ejecutar todas las cargas de trabajo para cada una de las bases de datos mencionadas. Y en el cuadro uno, los resultados numéricos obtenidos.

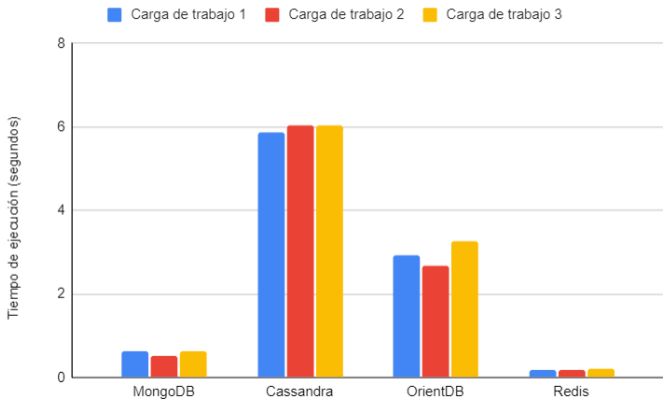

Figura 1: Resultados obtenidos de las las tres cargas de trabajo.

Al analizar los resultados, en una primera instancia se puede observar que la ejecución del tipo de operación sobre el volumen de datos no influye de manera decisiva en el tiempo de ejecución de la operación, es decir, tanto las lecturas como las escrituras tienen un tiempo relativo muy similar, siendo las escrituras más lentas, y por ende más costosas.

En el caso del trabajo original para una misma base de datos, los valores de lectura y escritura se diferencian más notoriamente. Por ejemplo, tanto la base de datos Redis como MongoDB se mantuvieron en valores muy cercanos, mientras que OrientDB y, especialmente, Cassandra dieron resultados equivalentes a un incremento cercano al 50 % y 1500 % respectivamente.

En el caso actual, las cuatro bases de datos presentaron resultados similares para los dos tipos de operaciones distintas.

Redis en particular, ha mantenido los resultados observados en ambos trabajos, aparte de ser la base que mejor ha respondido a las pruebas realizadas. Como se menciona en el trabajo original, las bases de datos de clave-valor son altamente optimizadas para dar buenos resultados a la hora de realizar las operaciones de lectura y escritura debido a que estructuralmente utilizan memoria volátil para mantener la referencias entre los registros almacenados.

MongoDB ha tenido una mejora sustancial en comparación al trabajo original. Si bien los valores absolutos no son comparables entre sí, los valores obtenidos en comparación a las otras bases de datos son notablemente mejores en los resultados actuales.

[1]https://docs.microsoft.com/en-us/windows/wsl/

Cuadro I: Resultados numéricos obtenidos de las distintas cargas de trabajo para las cuatro bases de datos expresados en segundos

	Carga de trabajo uno	***Carga de trabajo dos***	***Carga de trabajo tres***
MongoDB	0,628	0,503	0,63
Cassandra	6,047	6,04	5,878
OrientDB	3,254	2,673	2,922
Redis	0,207	0,185	0,18

OrientDB presentó resultados más similares a MongoDB, siendo una posible causa que ambas bases de datos están basadas en documentos. A su vez, los resultados relativos obtenidos de las tres cargas de trabajo también son similares entre sí. Del mismo modo que en el trabajo original, OrientDB resultó ser menos performante que MongoDB.

Por último, la base de datos Cassandra presentó resultados que llaman la atención y que dan indicios a una posible anomalía. En particular, en el trabajo original se menciona que las bases de datos basadas en familia de columnas son optimizadas especialmente para realizar operaciones de escritura, algo que se ve reflejado en sus resultados, ya que Cassandra retorna valores muy similares a Redis. En el caso actual, es varias veces más lenta que las otras opciones y a su vez muestra valores muy similares entre las operaciones de lectura y escritura. Una posible explicación para esto puede ser que la versión utilizada de Cassandra es la 4.0 que es la última versión RC[2] que se lanzó hace, aproximadamente, 20 días (2021-06-30).

VI. Conclusiones

En este trabajo se evaluaron cuatro de las base de datos no relacionales más populares[3] en la actualidad y pertenecientes a tres distintas categorías de base de datos, teniendo principal énfasis en las operaciones más comunes: lectura y escritura. Se utilizó la herramiena de Yahoo Cloud Serving Benchmark para la realización y ejecución de las pruebas. Los resultado que se obtuvieron fueron tales que la base de datos Redis, basada en clave-valor, obtuvo los mejores resultados en comparación a las otras bases de datos. Por otro lado, las bases de datos documentales presentaron un resultado similar pero relativamente por encima de los resultados de Redis. Y por último, Cassandra, presentó la peor performance entre las cuatros bases de datos aunque la versión que se utilizó no fue la última de la versión estable.

Por otro lado, también se puede concluir que los resultados absolutos en comparación al trabajo original, teniendo en cuenta la diferencia de utilizar una máquina con capacidades de computación actual y la última versión de todos los productos actual, son ampliamente mejores manteniendo la misma cantidad de datos y realizando el mismo tipo y cantidad de operaciones.

Por último, notar que las diferencias entre la realización de una operación de lectura o escritura son relativamente muy similares para las cuatro distintas bases de datos, una diferencia que sí se aprecia en el trabajo original.

[2]https://es.wikipedia.org/wiki/Ciclo_de_vida_del_lanzamiento_de_software#Versi\%C3\%B3n_candidata_a_definitiva_(RC)

[3]https://db-engines.com/en/ranking

VII. BIBLIOGRAFÍA

Which NoSQL Database? A Performance Overview - Veronika AbramovaJorge, BernardinoJorge Bernardino, Pedro FurtadoPedro Furtado