Desempeño de bases de datos no relacionales: MongoDB, ArangoDB, OrientDB y Elastic Search

Matilde Gómez de Salazar
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay
matilde.gomez@fing.edu.uy

***Resumen*—Se realizaron las mismas pruebas de rendimiento del artículo "Document Oriented NoSQL Databases: An Empirical Study" [Omji Mishra()], para comparar diferentes tipos de bases de datos NoSQL evaluar y estudiar diferencias y similitudes en los resultados.**

I. Introducción

Se realizaron pruebas utilizando la herramienta Yahoo! Cloud Serving Benchmark (YCSB) Con el objetivo de analizar y comparar rendimientos de operaciones en diferentes bases de datos documentales, para estudiar cuál es la más adecuada a utilizar según distintos factores como por ejemplo cantidad de clientes interactuando con el servidor, así como también la cantidad de consultas de lectura o escritura. Las pruebas son las mencionadas en el artículo "Document Oriented NoSQL Databases: An Empirical Study", dónde se carga una misma base de datos üsertable"para diferentes bases de datos NoSQL, la herramienta tiene diferentes "workloads"que son cargas de trabajo en diferentes proporciones de lectura y escritura.

Este documento se organiza de la siguiente manera, la sección dos presenta trabajos relacionados y una breve descripción de las cuatro bases de datos documentales que se utilizaron para comparar el rendimiento en las pruebas. En la sección tres se describe la preparación del ambiente y las características del equipo utilizado, la implementación de las pruebas, la ejecución y los resultados con su análisis. Por ultimo, en la cuatro se presenta las conclusiones.

II. Trabajos relacionados

Este informe toma como punto de partida tanto como Para la realización y comparación de las pruebas como para el análisis de los resultados el artículo "Document Oriented NoSQL Databases: An Empirical Study"de los autores Omji Mishra, Pooja Lodhi y Shikha Meht ya que éste realiza pruebas interesantes de rendimiento entre las bases de datos documentales MongoDB [MongoDB()], ArangoDB [ArangoDB()], OrientDB y Elastic Search.

MongoDB: Es una base de datos gratuita orientada a documentos de código abierto. Utiliza un formato de documento similar a JSON.

ArangoDB: Arango dB es una base de datos de código abierto multimodelo NoSQL. Ha sido diseñada especialmente para permitir que datos clavevalor, en documentos y grafos que puedan ser almacenados juntos, realizando consultas con un mismo lenguaje.También utiliza formato de documento de JSON.

OrientDB: Es una base de datos NoSQL escrita en JAVA. Es una base de datos multimodelo, por lo que a veces también se considera como una base de datos de grafos en lugar de datos orientados a documentos.

Elastic Search:Es un motor de búsqueda de código abierto basado en Lucene. Está desarrollado en JAVA.

III. Desarrollo

III-A. Características del equipo utilizado

Las pruebas se realizaron en un equipo con las siguientes características:

- Sistema Operativo: Microsoft Windows 10 Home
- Procesador: Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz 2.90 GHz
- Memoria: RAM 8,00 GB.
- Tipo de sistema: Sistema operativo de 64 bits, procesador basado en x64.

III-B. Implementación de las pruebas

Como se dijo anteriormente, Las pruebas se dividen en dos partes, por un lado, se va a analizar el tiempo de ejecución en segundos (runtime) y la cantidad de operaciones por segundo (throughput) para las diferentes cargas de operaciones (workloads) de lectura y escritura siguientes:

- Workload a : actualización pesada - 50 % lectura 50 % actualización
- Workload b : lectura pesada - 95 % lectura 5 % actualización
- Workload c : solo lectura - 100 % lectura
- Workload d : leer último - 95 % lectura 5 % incersión
- Workload f : leer, modificar, escribir - 100 % leer, modificar, escribir

Se utilizó el conjunto de datos üsertable"dado por la herramienta YCSB con mil registros. Por otro lado, se va comparar MongoDB y ArangoDB con el mismo conjunto de datos y cargas de operaciones pero para un millón de registros y las siguientes cantidades de hilos de cliente: 1, 3, 6, 10, 20, 32, 50, 64, 128.

Las pruebas constan de una parte de carga (load) y otra de ejecución (run), donde por ejeplo los comandos en consola para ejecutar una de las pruebas (workload a) de MongoDB es:

bin\ycsb load mongodb -P workloads\workloada -p recordcount=1000

bin\ ycsb run mongodb -P workloads\workloada -p recordcount=1000

III-C. Preparación del ambiente

Se instalaron en el sistema ArangoDB, ElasticSearch y MongoDB (para las pruebas con orientDB no fue necesario su instalación), así como también la herramienta de YCSB mencionada, Java sdk, Phyton y Maven.

III-D. Ejecución de las pruebas y problemas encontrados

Para ejecutar las pruebas se utilizó la guía que brinda [Yahoo(2015-2017)] para cada una de las bases de datos, para alguna de ellas fue suficiente con seguir la guia, pero para otras, se debío investigar algunos errores de dependencias en algún archivo al compilar o incluso falta de información de configuración las bases o de atributos para los comandos a ejecutar en consola.

III-E. Análisis de Resultados

En la primera etapa de pruebas los resultados estan reflejados en las siguientes gráficas:

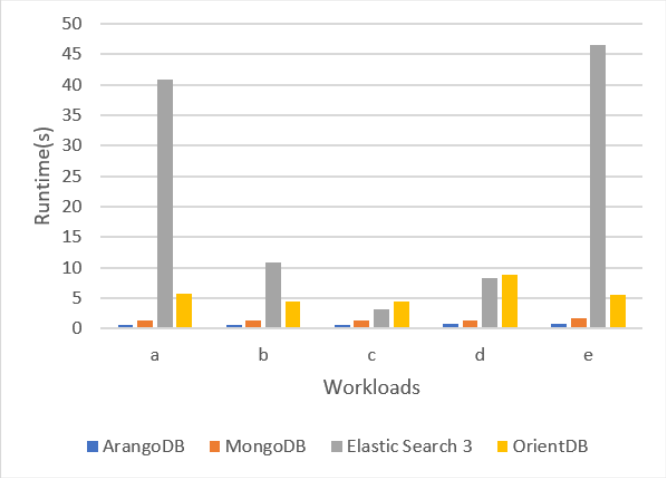

Figura 1: Runtime / workloads. Comparación diferentes bases de datos.

La figura 1 muestra notoriamente que al igual que en el artículo [Omji Mishra()], Elastic Search tiene el mayor tiempo de ejecución,por lo tanto no es apropiado utilizarlo en aplicaciones donde el tiempo de respuesta sea un factor importante, en particular las operaciones de actualización son las que más costo de tiempo tienen para Elastic Search. Los resultados indican a OrientDB como el segundo con peores resultados, pero con valores mucho más deseables que Elastic Search. Entre ArangoDB y MongoDb aparece la primera diferencia de resultados ya que en los resultados aquí indicados determinan a ArangoDB con los mejores valores en contraposición con los resultados del artículo.

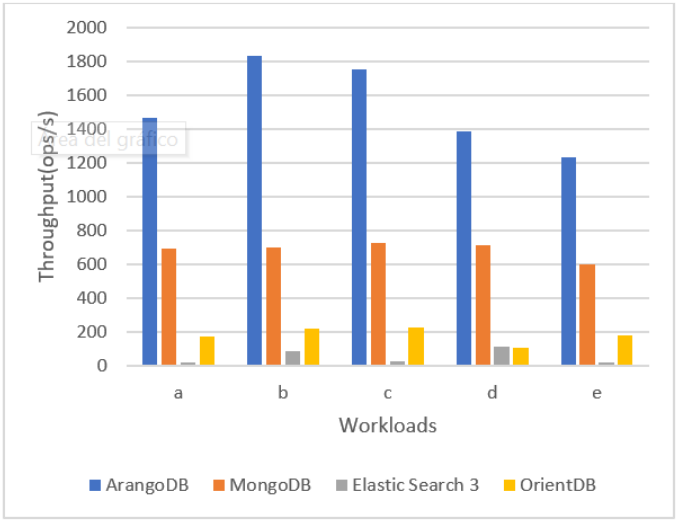

Figura 2: Throughput / workloads. Comparación diferentes bases de datos.

La figura 2 muestra también a Elastic Search como el peor en cantidad de operaciones por segundo, también muestra a OrientDB como el segundo peor, y existe también diferencias en los resultados en cuanto a señalar a ArangoDB como el que ejecuta más operaciones por segundo, duplicándolo en casi todos las cargas de trabajo a MongoDB, pero también ambas diferenciadas ampliamente del resto.

En la segunda etapa de pruebas se obtubieron los siguientes resultados:

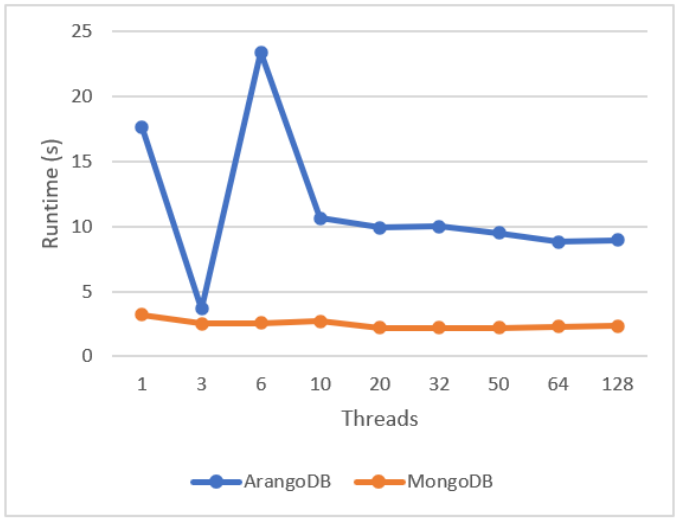

Figura 3: Runtime / Threads. Comparación MongoDB con ArangoDB, Workload A.

En la figura 3 coincide con el artículo en que MongoDB tiene un tiempo de respuesta bajo para todos los hilos en comaración a ArangoDB para grandes cantidades de actualizaciones, incluso todos los valores se mantienen por debajo de cinco segundos. Otra coincidencia es que en MongoDB para un hilo, los resultados fueron levemente mejores.

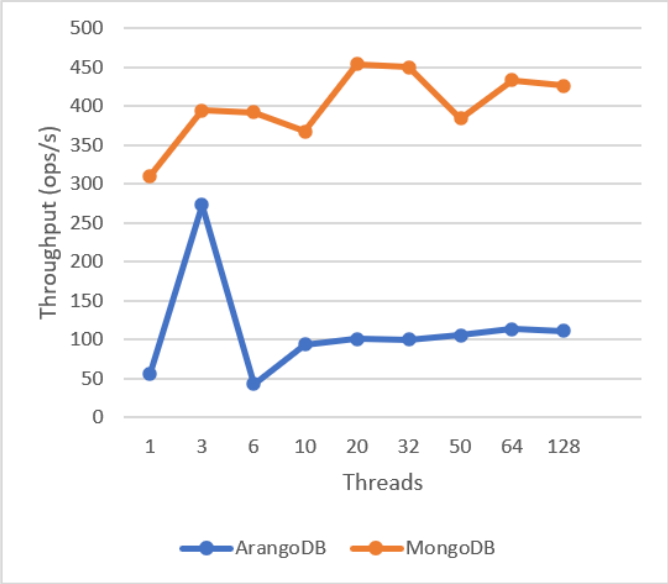

Figura 4: Throughput / Threads. Comparación MongoDB con ArangoDB, Workload A.

En la figura 4 también muestra que para un gran número de actualizaciones MongoDB es la mejor opción. El resultado de los valores es similar al artículo con la excepción del valor tres para ArangoDB que hace un salto creciente, lo que uno supone que influyó un factor externo.

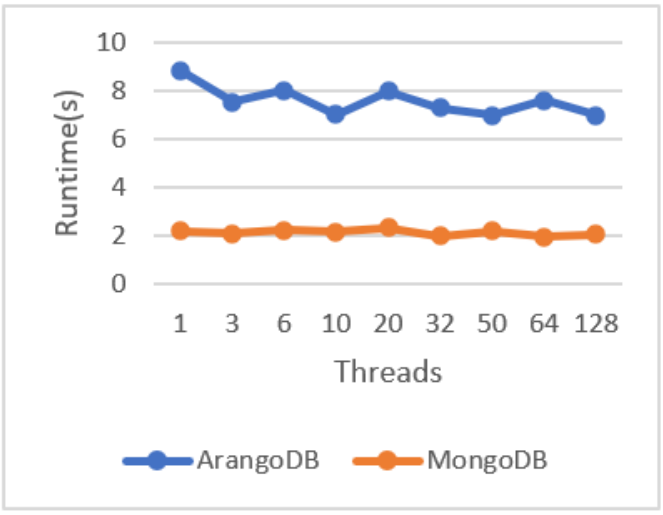

Figura 5: Runtime / Threads. Comparación MongoDB con ArangoDB, Workload B.

En la figura 5 también muestra que también para un gran número de lecturas MongoDB es la mejor opción independientemente de los hilos de clientes. Los resultados obtenidos son muy similares con MongoDB estando en el entorno de dos segundos en tiempo de ejecución, pero ArangoDB tiene un promedio de ocho segundos, mientras que en el artículo se encuentra en el entorno de los quince.

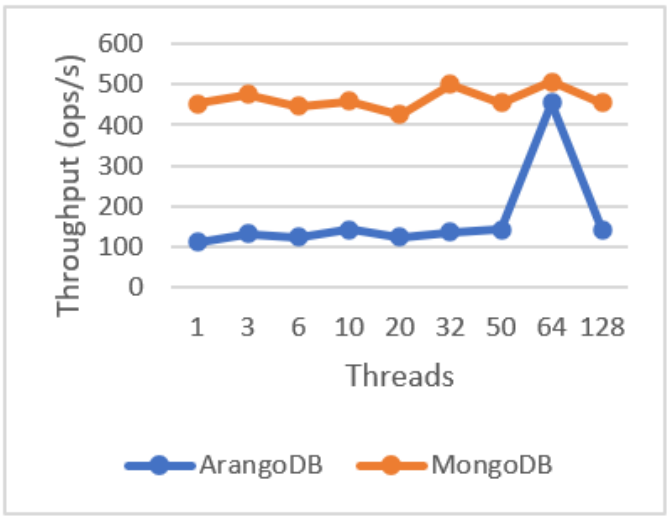

Figura 6: Throughput / Threads. Comparación MongoDB con ArangoDB, Workload B.

En la figura 6 también muestra a ArangoDB con peor desempeño de operaciones realizadas por segundo, y al igual que en la figura 4 existe un salto, pero en este caso, cuando la cantidad de hilos de clientes es 64.

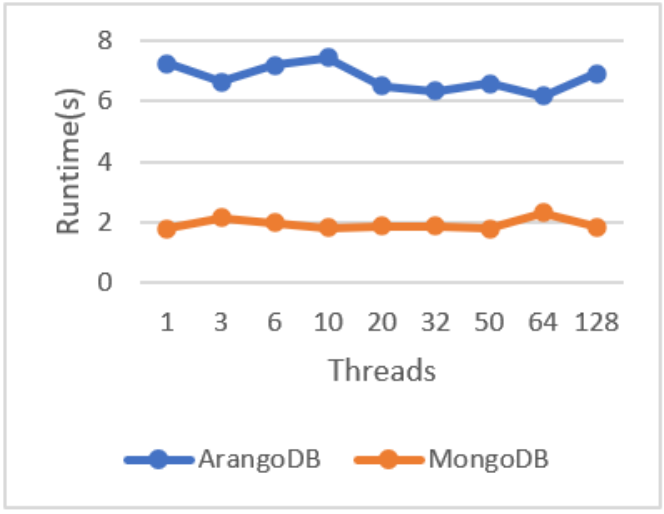

Figura 7: Runtime / Threads. Comparación MongoDB con ArangoDB, Workload C.

Para operaciones de solo lectura mostrados en las figuras 7 y 8 los resultados fueron muy similares a las operaciones de mayoría lectura (Workload B), donde muestra con mejor desempeño tanto en operaciones por segundo como en tiempo de respuesta, lo cual es coherente ya que en una carga de trabajo existe un 95 % de lecturas y en el otro, un 100 %.

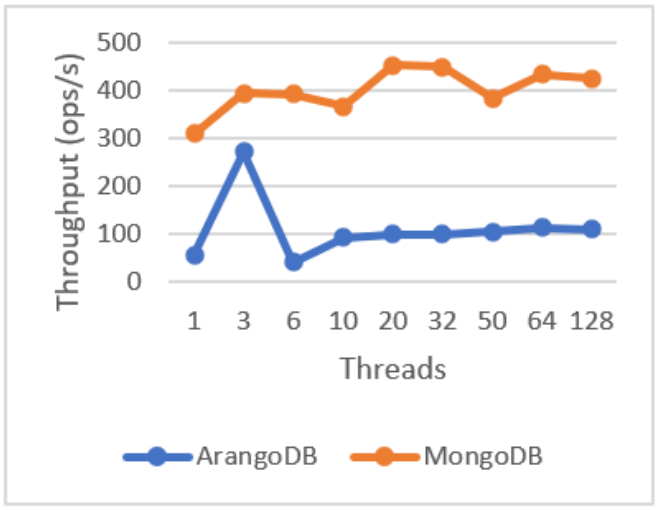

Figura 8: Throughput / Threads. Comparación MongoDB con ArangoDB, Workload C.

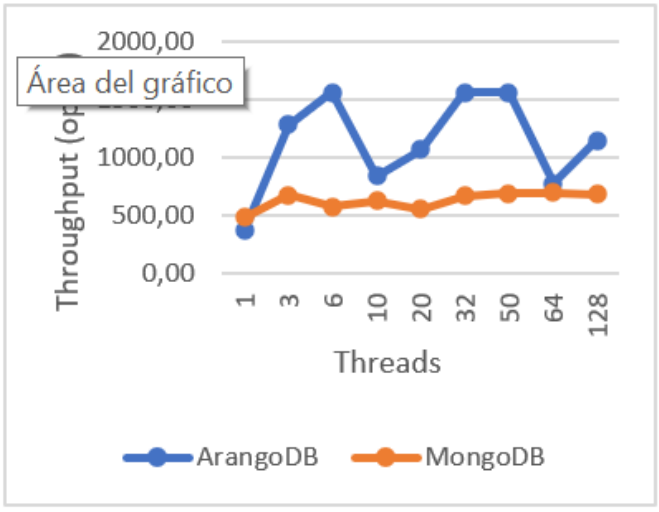

Figura 10: Throughput / Threads. Comparación MongoDB con ArangoDB, Workload D.

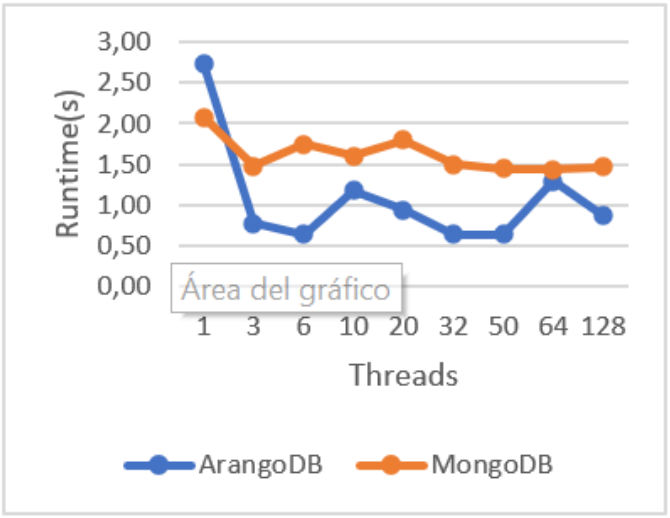

Figura 9: Runtime / Threads. Comparación MongoDB con ArangoDB, Workload D.

Para las pruebas de leer el más reciente y de leer, modificar y escribir donde los resultados se muestran en las figuras 9, 10, 11 y 12 los resultados dieron muy favorables para ambos modelos de bases de datos, esto implica valores muy similares en MongoDB tanto para el artículo [Omji Mishra()] como para las pruebas aquí mostradas, pero actualmente dieron también mejores valores para ArangoDB, incluso superando a MongoDB en algunos casos.

Esto es raro, pero al no haber una diferencia muy grande para estos casos particulares donde los resultados se mostraron intercambiados, los valores entre ambos modelos igual siguen indicando a MongoDB con mejor desempeño en el global de todas las pruebas.

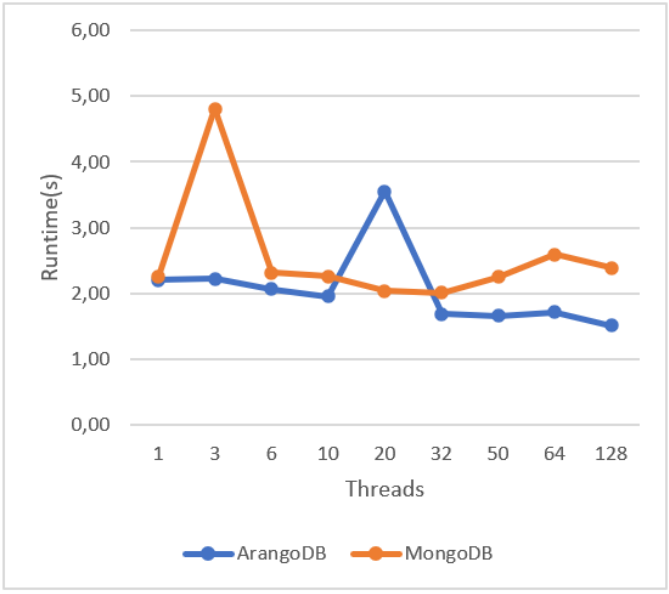

Figura 11: Runtime / Threads. Comparación MongoDB con ArangoDB, Workload F.

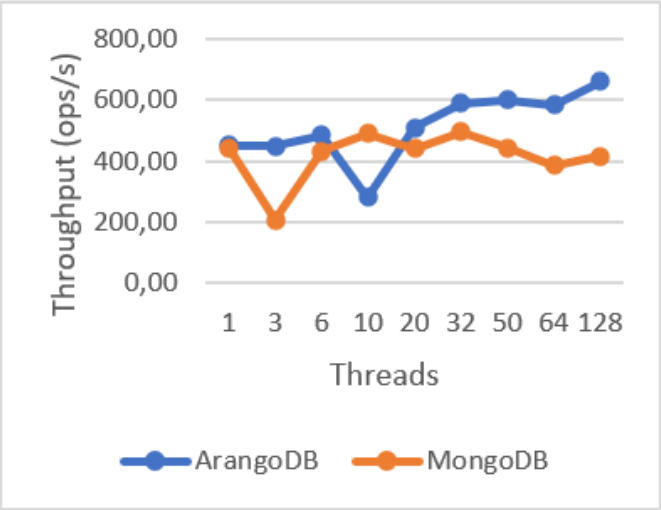

Figura 12: Throughput / Threads. Comparación MongoDB con ArangoDB, Workload F.

IV. Conclusiones y trabajo futuro

IV-A. Conclusiones y trabajo futuro

Se concluye que MongoDB tanto en el trabajo realizado en este estudio, así como también en el realizado por Pooja Lodhi y Shikha Mehta Omji Mishra, es el que obtuvo mejor despeño para todas las pruebas. Luego para la primera etapa Elastic Search fue el que obtuvo peores resultados, con valores significativamente peores al resto.

Algo que se podría haber realizado pero no se llegó por la falta de tiempo ya que al momento de realizar las pruebas tanto para ArangoDB como para OrientDB llevó varios días, que no estaban previstos de investigación en diferentes foros, algunos de los cuales eran los mismos que realizaron la herramienta de yahoo los que contestaban a problemas similares,para poder ejecutalas, ya que no se lograba la carga o la compilación del proyecto, fue ejecutar las mismas pruebas para una maquina virtual ubuntu, para también observar el comportamiento, especialmente en las pruebas que los resultados fueron diferentes.

Referencias

[ArangoDB()] Inc ArangoDB. ArangoDB. URL https://www.arangodb.com/.

[MongoDB()] Inc MongoDB. MongoDB. URL https://www.mongodb.com/try/download/community.

[Omji Mishra()] Pooja Lodhi y Shikha Mehta Omji Mishra. Document Oriented NoSQL Databases: An Empirical Study.

[Yahoo(2015-2017)] Yahoo. Yahoo! Cloud Serving Benchmark, 2015-2017. URL https://github.com/brianfrankcooper/YCSB.