RESEARCHERID
THOMSON REUTERS

[Kaur, 5(8): August 2018]
DOI- 10.5281/zenodo.1400573

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

PERFORMANCE EVALUATION OF VARIOUS NOSQL DATABASES USING YCSB ON BASIS OF CRUD OPERATIONS USING PARALLEL VIRTUAL MACHINES

Mandeep Kaur[1] & Er. Gurjit Singh Bhathal[2]

*1&2 Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India

ABSTRACT

The advancement in internet technology, results in generating a large amount of data daily. Traditional databases are not capable of storing such huge data, as major concern is assess time, which is increased. So, the NO SQL database are capable of storing such huge data, with good access complexity. The major databases existing in market are Mongo, Cassandra, Hbase databases. These are major stake holders in market. In this paper, the performance of various NoSQL databases is evaluated using parallel virtual machines, i.e. using the threads. The tasks are executed in parallel, not only in single thread environment. The outcomes shows that Cassandra DB is outperformer database, whereas Mongo DB is suitable for smaller datasets. Hbase is good performer in between both the databases

I. INTRODUCTION

Relational databases have been satisfying the data storage needs of the computing applications since the inception of databases in the 1980s. The stronghold of these databases lies in the fact that it guarantees ACID (Atomicity, Consistency, isolation, Durability) properties. As much as these databases have been evolved and sustained the tide of change, these have failed miserably in satisfying the data storage needs of the modern computing applications. Digital world is growing outrageously and humongous data is generated by the current applications especially the social media applications like Facebook, Twitter. The term 'Data' has become more complex than ever in terms of its variety (structured and unstructured), volume (terabytes to petabytes) and velocity (tremendous growth). This complexity has given rise to a new term called 'big data'. Such data is being produced in huge amounts that it can't be supported by conventional database systems like RDBMS.

To support such complex kind of data, new type of databases has come into picture called NoSQL databases. These refers to a class of non-relational databases that can cope with the large quantities of complex, ever increasing data which cannot be restricted into restricted into tables or relations. Since these are NoSQL databases, these certainly don't need SQL for data manipulation or querying the data. In order to induce large-scale data storage and to perform parallel operations across a large number of commodity servers, NoSQL databases are specially designed to be distributed, parallel, scalable and non-relational. In these databases, the full support of ACID properties as in relational databases is abandoned in order to attain horizontal scalability, parallelism and enhanced performance. Comparison of NoSQL databases and relational databases is shown in figure 1.

ISSN 2348 – 8034
Impact Factor- 5.070

Feature	NoSQL Databases	Relational Databases
Performance	High	Low
Reliability	Poor	Good
Availability	Good	Good
Consistency	Poor	Good
Data Storage	Optimized for huge data	Medium sized to large
Scalability	High	High (but more expensive)

Figure 1. Comparison of NoSQL and Relational databases. [1]

II. LITERATURE REVIEW

Abramova et.al. [2] carried out a detailed comparative analysis various NoSQL databases by making use of Yahoo! Cloud serving Benchmark [3]. For experimental evaluation, databases used were Redis, Cassandra, Hbase, MongoDB and OrientDB. 600,000 records were generated randomly and used with different loads by changing the ratios of insert, update and delete operations. Experimental results indicated that Redis database performed best of all and column family databases like Cassandra, Hbase are best suited for update operations.

Li et.al. [4] Compared the performance of various NoSQL databases like MongoDB, RavenDB, CouchDB, Cassandra, Hypertable, Couchbase and MS SQL Express. These databases were evaluated against five experiments: 1)Time to instantiate database bucket 2) Time to read values corresponding to given keys 3) time to write key-value pairs 4) Time to delete key-value pairs 5) Time to fetch all keys. The data for which experiments were conducted ranged from 10 records to 100,000 records. Results indicated that Couchbase and MongoDB performed tremendously better than other in all kinds of read insert and delete operations.

Boicea et.al [5] carried out the comparative analysis of MongoDB and Oracle in order to have a clear view of the performance difference between SQL and NoSQL databases. The databases are evaluated against three experiments: 1) Time elapsed to insert the data 2) Time elapsed to update the data 3) Time elapsed to delete the data. Experimental results exhibited MongoDB as the clear winner in all the operations. Records were varied from 10 to 1000, 00.

Konstantinou et.al. [6] came forth with a comparison of Cassandra, HBase and Riak in terms of read and update operations in order to perform a study regarding the elasticity of non-relational databases. The study concludes that Cassandra is best suited for write operations and Hbase has the higher elasticity and better suited for read operations. On the other hand, Riak does not exhibit any increase in performance irrespective of any operations.

Van der Veen et.al. [7] Compared Cassandra, MongoDB and PostgreSQL in order to find out the database which performs best in single server and distributed servers scenario. The results conclude that Cassandra performs best in a distributed server scenario while MongoDB gives high throughput in a single server scenario.

Nelubin et.al. [8] tested the failover characteristics of different NoSQL databases like Aerospike, Cassandra, Couchbase and MongoDB. Results showed that MongoDB has the least favourable downtime, then Cassandra and Aerospike has the lowest downtime. The limitation of this study was that it is not suited for real-world scenarios because the datasets used for conducting the study was RAM only datasets. The study further concludes that MongoDB was not suited for highly available system.

ISSN 2348 – 8034
Impact Factor- 5.070

III. PROPOSED WORK

In this paper the major three databases are compared on basis of certain parameters. The MongoDB, CassandraDB and HBase are evaluated. The machine used to compute the performance having configuration 8 GB DDR4, 2 TB hard disk, and Nvidia 1050 Graphic card with i7 processor. The Hadoop user is created. In Ubuntu 14. 04 with 2.0. 6 version of HBase. The Git for YCSB is cloned to our system. Various tools are also required to see the database like robo mongo is used for MongoDB and Graphical user interface is available for Hbase. The Java environment with java virtual machine is installed to system.

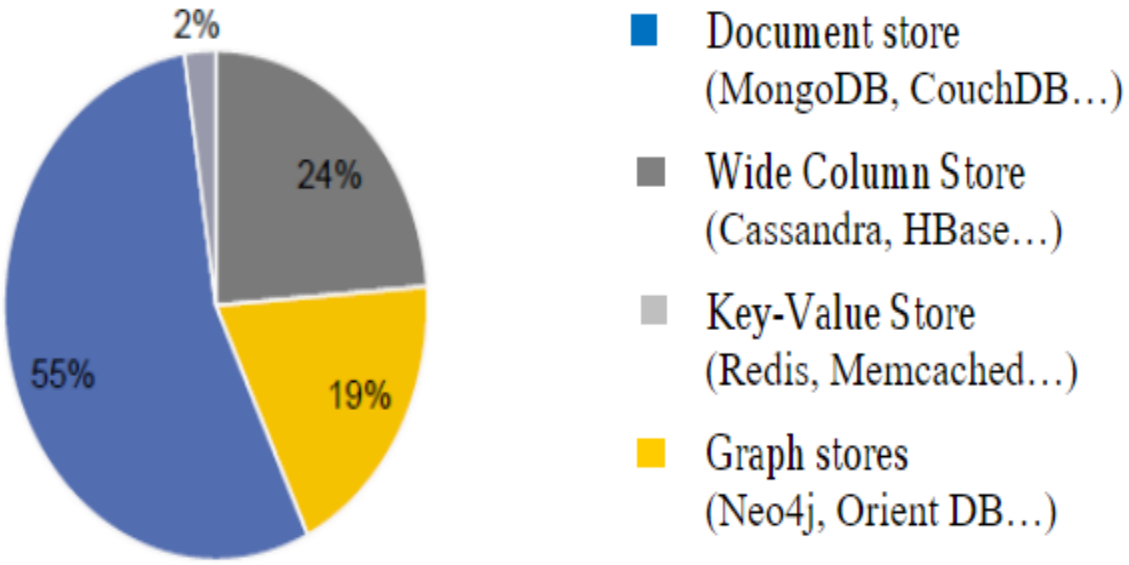

Figure2. Most used databases

a) The YCSB is cloned to system. (git clone https://github.com/ycsb)
b) Installation of all the required databases with setting up of java environment.
c) Maven is automation tool, which use to fetch data from various APIs.
d) Various workloads are given in YCSB folder, which are executed using suitable slf4j files.
e) The performance of databases is evaluated on bases of CRUD operations.

The three workloads are used workload A (100% Insert Operations), workload B (100% Read Operations) and workload C (100% Update Operation). The workload provided is 1000, 20000, 40000, 60000, 80000, 100000 operations for all the cases. All the database are running 10 threads each and operation are executed in parallel.

IV. RESULTS

The results for all the databases are evaluated on basis of insert, read and update operations. The execution time is evaluated for all the databases. Lesser the execution time, the performance of database is better. The major cases to evaluate the performance are taken as:

- Insert Operations
- Update Operations
- Read Operations

All the databases are evaluated using different datasets starting from 10000 to 100000. The operations are executed in parallel by threads. The resource dependencies are existing among threads.

ISSN 2348 – 8034
Impact Factor- 5.070

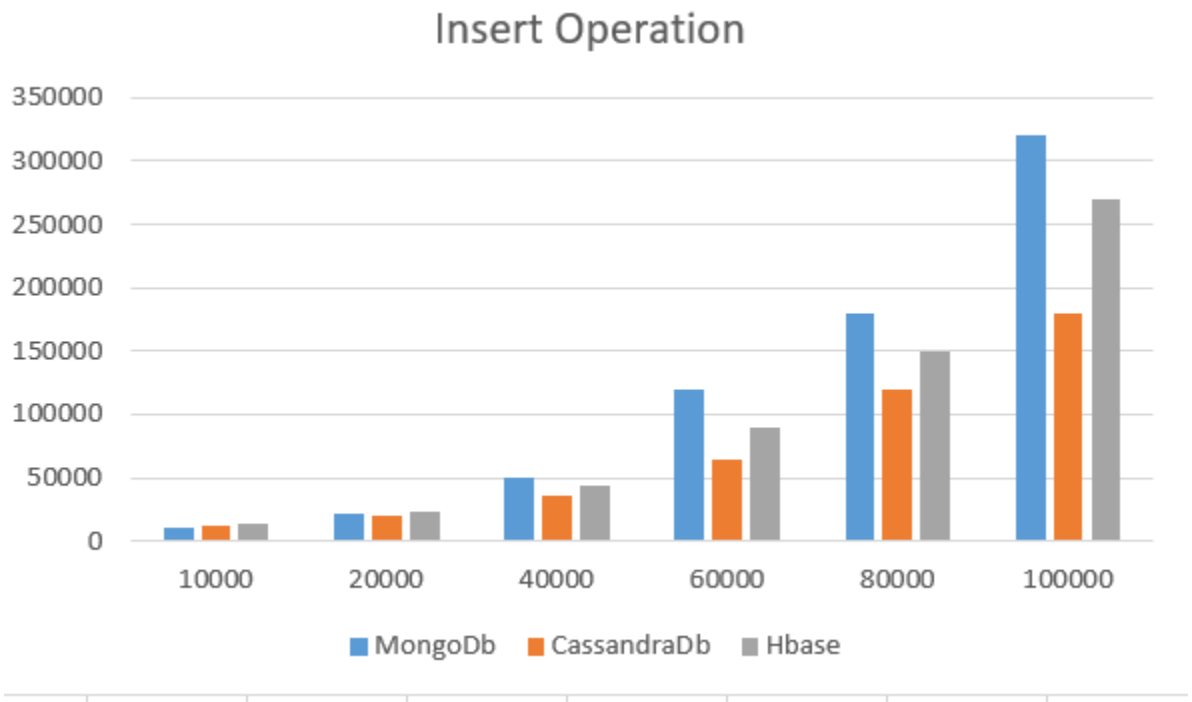

Figure 3. Time latency for insert operation in milliseconds

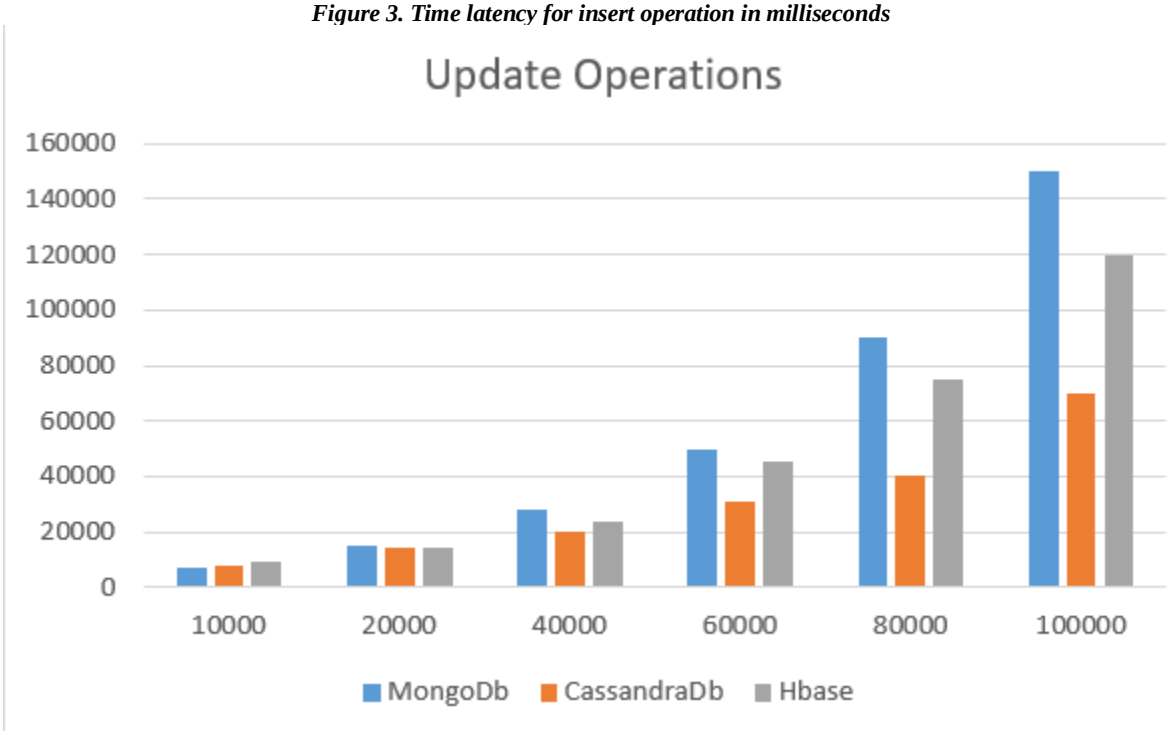

Figure 4. Time latency for update operation in milliseconds

ISSN 2348 – 8034
Impact Factor- 5.070

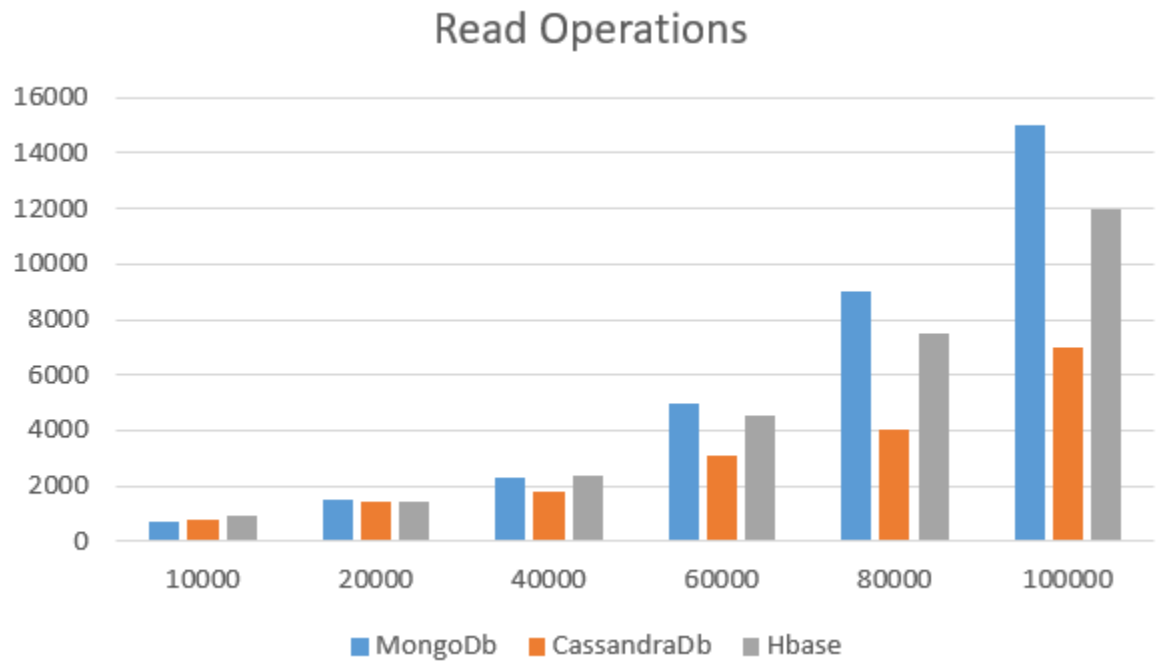

Figure 3. Time latency for read operation in milliseconds

The Cassandra outperform in terms of all the operations. The MongoDB is suitable, when data to store is small. Hbase is also a good performer, its performance is lesser in compare to Cassandra, but higher in security.

V. CONCLUSION

No-SQL databases are almost used in every field, and they are capable to store huge amount of data. They are highly secured in contrast with various SQL or traditional databases, and there performance is very good. In this paper, the three major databases are compared and evaluated on basis of certain parameters like insert, update and read. The Cassandra is found to be beast candidate almost for all the cases. The Mongo DB is quite suitable for smaller size databases, but as data size increases the performance also degraded. The scalability factor is very poor for Mongo Db but Cassandra Db performance improves with add on in data. Hbase is also a good competitor in market, as performance of Hbase is better than Mongo Db and security and scalability is far better than Mongo Db

REFERENCES

1. https://blog.pandorafms.org/nosql-databases-the-definitive-guide/
2. *Abramova, V., Bernardino, J., & Furtado, P. (2014). Which nosql database? a performance overview. Open Journal of Databases (OJDB), 1(2), 17-24.*
3. *Cooper, B. F., Silberstein, A., Tam, E., Ramakrishnan, R., & Sears, R. (2010, June). Benchmarking cloud serving systems with YCSB. In Proceedings of the 1st ACM symposium on Cloud computing (pp. 143-154). ACM.*
4. *Li, Y., & Manoharan, S. (2013, August). A performance comparison of SQL and NoSQL databases. In Communications, computers and signal processing (PACRIM), 2013 IEEE pacific rim conference on (pp. 15-19). IEEE.*
5. *Boicea, A., Radulescu, F., & Agapin, L. I. (2012, September). MongoDB vs Oracle--database comparison. In 2012 third international conference on emerging intelligent data and web technologies (pp. 330-335). IEEE.*

6. *Konstantinou, I., Angelou, E., Boumpouka, C., Tsoumakos, D., & Koziris, N. (2011, October). On the elasticity of NoSQL databases over cloud management platforms. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 2385-2388). ACM.*
7. *Van der Veen, J. S., Van der Waaij, B., & Meijer, R. J. (2012, June). Sensor data storage performance: SQL or NoSQL, physical or virtual. In Cloud computing (CLOUD), 2012 IEEE 5th international conference on (pp. 431-438). IEEE.*
8. *Nelubin, D., & Engber, B. (2013). Nosql failover characteristics: Aerospike, Cassandra, Couchbase, MongoDB. Thumbtack Technology.*