ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 3; Issue 6; November-December-2015; Page No. 37-40

A Study of YCSB –tool for measuring a performance of NOSQL databases

Mrs. Rohini Gaikwad[1], Dr. A. C. Goje[2]

[1] Research Scholar, Vidya Pratishthan's Institute of Information Technology, Savitribai Phule Pune University, Baramati, Pune, Maharashtra, 413133.

[2] Research Guide, Vidya Pratishthan's Institute of Information Technology, Savitribai Phule Pune University, Baramati, Pune, Maharashtra, 413133.

Abstract

Day by day the massive amount of data increased through various sources in various formats. Mostly the semi-structured and unstructured data has been generated, which causes the performance of RDBMS. This is generated the need of special databases viz NOSQL. Each Nosql database has its own strength and weaknesses; hence the business organization has facing the problem while selection .There is no standard framework which can suggest the selection of database system. Therefore, the performance is an important factor for deciding which database will be used for enterprises and applications. Therefore, it is necessary to compare and analyze the execution time of difference NoSQL databases, and provide a performance reference.

Currently, there are more than 150 NoSQL databases with diverse features and optimizations [1], and a number of NoSQL databases provide all new features and advantages while keeping data consistent or even eventually consistent, depending on the system needs.

The YCSB plays an important role in comparison of various NOSQL databases performance evaluation.

The aim of this paper to study of YCSB benchmark tool for comparing performance of NOSQL database. This paper mainly focuses on literature review. This work will help the academic researcher those are working on NOSQL database performance issues for their work**.**

Keyword: YCSB, Workload, NOSQL Database, performance

1. Introduction

Now a day lots of work on NOSQL databases testing is carried out. There is a variety of papers, such as [12,13,14], which given overall analysis and presented theoretical approaches to describing characteristics and mechanisms of NoSQL databases.

Still there no standard framework for selection of NOSQL databases. Due to increased interests in NoSQL databases have been analyzed from application perspective, organization need. Therefore, the research of their performance, characteristics and used mechanisms, has been increased.

The performance is an important factor for deciding which database will be used for organization and their applications.

There are other benchmarks available, such as, TPC-H or SSB, which could be used to evaluate database performance, but the YCSB is most popular because of its simplicity.

The rest of the paper is organized with what is YCSB tool followed by literature review of research papers, articles followed by Findings and then conclusion.

2] What is YCSB

Yahoo Cloud Service Benchmark Client. A key design goal of this tool is extensibility as it can be used to benchmark new cloud database systems. This tool is available under an open source license. It has ready adapters for different NoSQL Databases. YCSB tool allows benchmarking multiple systems and comparing them by creating "workloads". Using this tool, one can install multiple systems on the same hardware configuration, and run the same workloads against each system.

Corresponding author: Mrs. Rohini Gaikwad

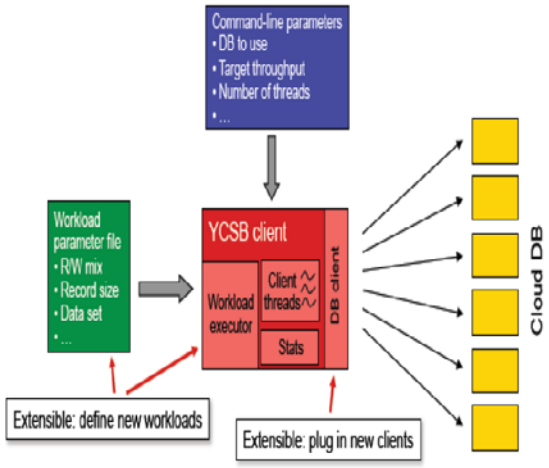

Fig 1: Architecture of YCSB [4][6]

3] Workload:

[3][22]The YCSB framework contains a core set of workloads to evaluate different aspects of a system's performance, called the YCSB Core Package. In YCSB, a package is a collection of related workloads. The workload defines the data that will be loaded into the database during the loading phase, and the operations that will be executed against the data set during the transaction phase, and can be used to evaluate systems at one particular point in the performance space. While the core package examines several interesting performance axes, YCSB have not attempted to exhaustively examine the entire performance space. It is developed in such a way that users of can develop their own packages either by defining a new set of workload parameters, or if necessary by writing Java code.

Operations within workload

[22][15]Operations against a data store were randomly selected and could be of the following types:[23]

1. Insert: Inserts a new record.
2. Update: Updates a record by replacing the value of one field.
3. Read: Reads a record, either one randomly selected field, or all fields.
4. Scan: Scans records in order, starting at a randomly selected record key. The number of records to scan is also selected randomly from the range between 1 and 100.

The defined workloads are:

1. Workload A: Update heavily
2. Workload B: Read mostly
3. Workload C: Read only
4. Workload D: Read latest
5. Workload E: Scan short ranges
6. Workload F: Read-modify-write
7. Workload G: Write heavily

Each workload is defined by:[15][19][22]

1. The number of records manipulated (read or written)
2. The number of columns per each record
3. The total size of a record or the size of each column
4. The number of threads used to load the system

4] Literature Review:

In [3], researchers presented performance comparison of NOSQL database using YCSB benchmark tool where the resources are limited. They have used only single PC for their experiment. Also the define a core set of benchmarks and generate a result for MongoDB, Elastic Search, Redis, and Orient DB implementation.

[3][4][5] Presents the YCSB framework to assist performance comparisons of the new generation of cloud data serving systems. They define a foundation set of benchmarks and report results for four widely used systems: Cassandra, HBase, and Yahoo!'s PNUTS, and a simple sharded MySQL implementation. Their work do not deal with a situation where resources are limited and do not reflect on some popular NoSQL databases like Mongodb, Redis, e.tc in their experiment.

[7] In this paper, researchers tried to make the best comparisons possible based on architectural arguments alone. They recommended getting some useful objective data comparing the architectures. The focus of this paper was on Horizontal Scalibility and Simple Operations

[9] Researchers compared NoSQL databases MongoDB and Hive with SQL Server PDW using YCSB [5] and TPC- H DSS [3] benchmarks. They compare these technologies for data analysis and interactive data serving. They concluded that though relational databases perform better, NoSQL systems have its own advantages such as elastic data models, auto-sharding and load balancing.[11]

[10] compares databases in terms of query performance, based on reads and updates, taking

into consideration Workload A, Workload C and Workload H. They have not considered the all workloads of YCSB benchmark in this paper.

Table1: Executed Workload [10]

Workload	%Read	% 1Write
A	50	50
C	100	0
H	0	100

[16]In this paper, authors presented an evaluation of MongoDB, Cassandra, HBase and SciDB using YCSB and two scientific data sets. Their results indicate that careful understanding of the distribution of the workload as well as aspects such as client side tools and parameters need to be considered to get the optimal performance from these databases.
[17] Presented experience and a comprehensive performance evaluation of six modern data stores. They also presented the benchmarking effort on e key-value stores. They compare the throughput of Apache Cassandra, Apache HBase, Project Voldemort, Redis, VoltDB, and a MySQL Cluster. The comparison is restricted to Key value stores only.

5] Findings

1] Due to the heavy mixed type data generation the need of NOSQL arises. But the lack of standard framework, the database selection is in dilemma.
2] The performance of database system standards on various aspects like Feature trade-off, Performance trade-off or any other.
3] YCSB is a pack of the *YCSB* Client, an extensible workload generator. It provides core workload as well facility to add our own workload
4] It used to compare relative performance of NoSQL database management systems focusing on throughput and latency.
5] While installation & configuration one has to take care of the versions of java, maven ,Github and Python. Version incompatibility will create the hurdle while building the project.
6] YCSB doesn't test for correctness.

6] Conclusion:

In this paper researchers studied various research papers, articles and white papers. In this paper we studied the need of YCSB tool its components, the terminology of workloads. The work of various researchers proves that the YCSB is the best tool for measuring the performance of NOSQL database in distributed environment.

The current versions of YCSB mainly work for scalability, throughput and latency. Now the new version arises for replication task.

Due to the change in nosql databases version, accordingly the need of changes in YCSB is also required and it is already running.

If a package of YCSB client, java and Maven bind together and make available, then it helps to reduce the problem occurring in the installation and building project and YCSB become more user friendly.

Acknowledgement:

I would like to thanks to all researchers and authors whose work in helped me to proceed in my work.

7] References:

1. NoSQL - http://nosql-database.org/.
2. The TPC-H Benchmark. http://www.tpc.org/tpch/
3. Yusuf Abubakar, Department of Computer Science, Nuhu Bamalli Polytechnic, Zaria - Nigeria S. Adeyi ,Department of Mathematics ,Ahmadu bello University,Zaria-Nigeria Ibrahim Gambo Auta ,Waziri Umaru Federal Polytechnic: "Performance Evaluation of NoSQL Systems Using YCSB in a resource Austere Environment" International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 September 2014 – www.ijais.org
4. B. F. Cooper et al. PNUTS: Yahoo!'s hosted data serving platform. In VLDB, 2008.
5. B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears, "Benchmarking cloud serving systems with YCSB," in Proceedings of the 1st ACM symposium on Cloud computing, ser. SoCC '10. New York, NY, USA: ACM, 2010, pp. 143–154. [Online]. Available: http://doi.acm.org/10.1145/1807128.1807152
6. Brian F. Cooper: Yahoo! Cloud Serving Benchmark, Overview and results – March 31, 2010 cooperb@yahoo-inc.com. Joint work with Adam Silberstein, Erwin Tam, Raghu Ramakrishnan and Russell Sears System setup and tuning assistance from members of the Cassandra and HBase committers, and the Sherpa engineering team.

7. Rick Cattell ,Originally published in 2010, last revised December 2011: “Scalable SQL and NoSQL Data Stores”
8. P. Shivam et al. Cutting corners: Workbench automation for server benchmarking. In Proc. USENIX Annual Technical Conference, 2008.
9. A. Floratou, N. Teletia, D. Dewitt, J. Patel, and D. Z. Zhang. “Can the elephants handle the nosql onslaught?” VLDB, 2012.
10. ABRAMOVA, Veronika; BERNARDINO, Jorge; FURTADO, Pedro - Which NoSQL Database? A Performance Overview. "Open Journal of Databases". ISSN 2199-3459. Vol. 1 Nº. 2 (2014) p. 17-24
11. E. Dede, M. Govindaraju SUNY Binghamton Binghamton,D. Gunter, R. Canon, L. Ramakrishnan Lawrence Berekely National Lab Berkeley, CA {dkgunter, scanon, lramakrishnan}@lbl.gov: “Performance Evaluation of a MongoDB and Hadoop Platform for Scientific Data Analysis”
12. Hecht, R. and JABLINSKI, S.: NoSQL Evaluation A Use Case Oriented Survey. Proceedings International Conference on Cloud and Service Computing, pp. 12-14, 2011. *Open Journal of Databases*
13. Han, J.: Survey on NOSQL Databases. Proceedings 6th International Conference on Pervasive Computing and Applications, pp. 363-366, 2011.
14. Leavitt, N.: Will NoSQL Databases Live up to Their Promise?, Computer Magazine, 43(2): 12-14, 2010.
15. Prasoon Kumar,MongoDB #CMGIndia:NoSQL Database Benchmarking
16. Lavanya Ramakrishnan, Pradeep K. Mantha, Yushu Yao, Richard S. Canon Lawrence Berkeley National Lab Berkeley, CA [lramakrishnan,pkmantha,yyao,scanon]@lbl.gov: Evaluation of NoSQL and Array Databases for Scientific Applications
17. Tilmann Rabl, Mohammad Sadogh, HansArno Jacobsen, Middleware Systems Research Group ,University of Toronto, Canada: Solving Big Data Challenges for Enterprise application Performance Management
18. https://github.com/brianfrankcooper/YCSB
19. Sergey Bushik, senior R&D engineer at Altoros Systems Inc:A vendor-independent comparison of NoSQL databases: Cassandra, HBase, MongoDB, Riak 22.10.2012 kl 20:53 | Network World (US)
20. Datastax corporation: Benchmarking top NOSQL databases,A performance comparison for architects and IT managers, February 2013.
21. Impetus, Innovation Architected: TPC-H for NOSQL performance benchmark
22. http://dl.acm.org/citation.cfm?id=1807152
23. http://www.dbms2.com/2013/01/17/ycsb-benchmark-notes