The European Proceedings of
Social & Behavioural Sciences
EpSBS

Future Academy ISSN: 2357-1330

http://dx.doi.org/10.15405/epsbs.2017.12.26

ICPE 2017
International Conference on Psychology and Education

EVALUATION OF NOSQL DBMS SCALABILITY FOR STORING DATA FOR PSYCHOLOGY WEB PLATFORM

Evgeny Nikulchev (a)*, Dmitry Ilin (b), Pavel Kolyasnikov (c), Sergey Kasatonov (d), Ilya Zakharov (e)
*Corresponding author

(a) Moscow Technological Institute, 38A, Leninskiy pr., Moscow, Russia & Russian Academy of Education, 8 Pogodinskay str., Moscow, Russia, 119435, nikulchev@mail.ru
(b) Moscow Technological University, 78, Vernadskogo prospect, Moscow, Russia & Moscow Technological Institute, 38A, Leninskiy pr., Moscow, Russia, i@dmitryilin.com
(c) Moscow Technological Institute, 38A, Leninskiy pr., Moscow, Russia, pavelkolyasnikov@gmail.com
(d) Moscow Technological University, 78, Vernadskogo prospect, Moscow, Russia, Russia, kasatonovsergey@yandex.ru
(e) Psychological Institute of Russian Academy of Education, 9-4, Mokhovaya str., Moscow, Russia, iliazaharov@gmail.com

Abstract

The study examines the design and development of a web-based platform for conducting online psychological research. The authors substantiated the need for the development of a platform and described advantages over the pen-and-paper approach, both with respect to data collection and subsequent processing. The necessity of using NoSQL solutions for data storage is indicated, which is due to the large number of users of the system and the need to work with large volumes of data arriving at high speed. As the main advantage the possibility of sharding is considered, which is offered by NoSQL solutions by default. Replication is considered as an additional method for improvement of the platform availability and fault tolerance. An experiment was performed to analyze the scalability of NoSQL database management systems MongoDB and Cassandra, which estimated the speed of data recording with increasing the number of nodes (from 1 to 3 nodes). The performance gain is noticeably better with an increase in the write volume to about 10K per query for MongoDB and 100K per query for Cassandra. In view of Cassandra's high requirements for the qualifications of engineers, MongoDB was chosen as a solution for developing a web-based platform for conducting psychological research.

Keywords: Psychology research, web-based platform, software architecture, NoSQL DBMS, sharding, replication.

http://dx.doi.org/10.15405/epsbs.2017.12.26
Corresponding Author: Evgeny Nikulchev
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

1. Introduction

One of the trends in modern cognitive science is the slow but consistent movement towards computerized and automated tools for research and data analysis (Ismatullina, Zakharov, Nikulchev, & Malykh, 2016). There are still some areas of research where traditional pen-and-paper approach may be more suitable (e.g. in some clinical populations or in rural areas where it is hard to get access to internet or even computer), but in general computer-based methods has proven to be adequate to such fields as behavioral genetics (Ismatullina & Voronin, 2017; Voronin, Ismatullina, Zaharov, & Malykh, 2016; Zakharov, Ismatullina, & Voronin, 2016; Rimfeld, et al., 2017), neuropsychology (Luciana, 2003), developmental psychology (Tikhomirova & Malykh, 2016), cross-cultural studies (Verbitskaya, Zinchenko, Malykh, & Tikhomirova, 2017) and others. Advantages of the computer-based tools include more convenient way to organize and store databases, standardized test administration or automated response recording. It also allows the researchers to implement the artificial intelligent algorithms of data analysis, data management and output both for researchers and their participants.

Over the past decades, psychological research is gradually moving from laboratories to the Internet. The previous work (Zakharov, Nikulchev, Ilin, & Ismatullina, 2017) examined the existing web technologies for conducting psychological research and their advantages:

- Easy access via the Internet.
- More people who can participate in the study.
- Individualization of results and feedback for participants.
- Use of machine learning and artificial intelligence to process the data.

In addition, in the study (Zakharov, Nikulchev, Ilin, & Ismatullina, 2017) it was told about the beginning of the development of a web-based tool for psychological research that aims to diagnose psychological, psycho-physiological and cognitive evaluation. The tool should include a wide range of generally accepted psychological methods and the ability to include new ones that are presented by independent researchers.

Within the scope of this work, the necessity of applying NoSQL solutions (Han, Haihong, Guan, & Jian, 2011; Nikulchev et al, 2015) will be substantiated and the experiment on the scalability analysis of MongoDB (Abramova & Bernardino, 2013; Dede, Govindaraju, Gunter, Canon, & Ramakrishnan, 2013) and Cassandra (Abramova & Bernardino, 2013; Lakshman & Malik, 2010) will be carried out.

2. Problem Statement

It is expected that the platform will be used by a large number of people, and a significant part of them will actively use it for a limited period of time – the beginning of the school year in schools. According to rough estimates, the number of schoolchildren in Russia is about 14 million (The number of pupils and personnel of educational institutions of the Russian Federation (The forecast up to 2020 year and score trends up to the year 2030), 2013). It is easy to imagine that significant resources will be required to process incoming data on the results of psychological experiments. At the initial stage, only part of schools will participate in the experiment, which does not eliminate the need for horizontal scaling. Proceeding from what has been said, in addition to the time of access to data, the speed of data recording and storage volumes will be important.

There are several scaling options that are used in the DBMS (Gorton & Klein, 2014; Gudivada, Rao, & Raghavan, 2014; Corbellini, Mateos, Zunino, Godoy, & Schiaffino, 2017), but in this case sharding is required. Separate replication will not allow you to achieve a high write speed, but only provide quick access. It is possible to combine sharding and replication to improve both read/write performance and to ensure reliable data storage. In the previous work (Zakharov, Nikulchev, Ilin, & Ismatullina, 2017) it has already been said about the need for database replication.

In order to save human resources for maintaining the infrastructure, it is advisable to consider non-relational DBMSs offering the functionality of sharding out of the box. For consideration two examples of such DBMSs are taken:

- MongoDB, as one of the most common Document-Oriented Database Management Systems.
- Cassandra, also a known database, but from the Wide Column Store category.

Both DBMSs are open source software and are freely available, which allows them to be used to solve the task. However, it is needed to conduct an experiment and evaluate the scalability of both databases in order to determine which one will be most suitable for implementing the web platform for psychological research.

In addition to scaling the algorithmic part, data storage must also be scalable. The best approach for the project is the combination of sharding and replication. Sharding can provide a system with high I/O performance, while replication can help to ensure the availability of the service (Zakharov, Nikulchev, Ilin, & Ismatullina, 2017).

3. Research Questions

The following issues will be considered in the paper:

- Which of the above DBMSs is best suited for storing large amounts of data in the case of a web platform
- Which of the above DBMSs makes lesser requirements for the qualification of the system administrator?

4. Purpose of the Study

The purpose of this paper is to conduct an experiment to test and compare the NoSQL databases MongoDB and Cassandra to select the most suitable solution for the Web-based Platform for Psychology Research.

5. Research Methods

To determine the feasibility of using the MongoDB and Cassandra databases, an experiment was conducted, during which the following criteria were investigated for each DBMS under consideration:

- Peak load resistance for data recording.
- Horizontal scalability.

- Setup and support complexity.

All experiments were carried out on the same machine with the following configuration:

- Processor: Intel Core i5-3570K 3.4 GHz.
- Memory: 8 GB 1333 MHz
- Internal drive: Intel SSD 520 Series.
- Operating system: Windows 7 Ultimate SP1.

To test the DBMS, virtual machines were created running OS Linux. In order to create similar machines with automatic system configuration, the Vagrant (Stillwell & Coutinho, 2015) wrapper was used. The system used Ubuntu/trusty64 box, version 20170202.1.0.

For MongoDB there were created:

- 3 shards (mongod): 1 CPU, 1536 MB RAM, speed limit for writing to the drive is 2 MB/s.
- 1 server (mongos): 1 CPU, 512 MB RAM.
- 1 client: 1 CPU, 512 MB RAM.

For Cassandra there were created:

- 3 shards: 1 CPU, 2048 MB RAM, speed limit for writing to the drive is 2 MB/s.
- 1 client: 1 CPU, 512 MB RAM.

The sizes of RAM are determined by the minimal requirements of a DBMS to machines.

In the course of the experiment, the client generates a random string of characters of a given length and continuously sends requests to write it to the database. During this we record the readings about the load on the shard processors and mongos, as well as the amount of RAM used.

For each tested DBMS, three series of four experiments were produced. Each series used a different number of shards: 1, 2 and 3 shards for 1st, 2nd and 3rd series respectively. Within a single series, experiments differ from each other in the number and size of records, but the total amount of recorded information remains approximately the same.

6. Findings

The results of the experiments are presented in Tables through 6. In the tables for each virtual machine there are displayed:

- CPU utilization in percent.
- Load on RAM (RAM usage) in megabytes; the number after the slash is the total amount of memory on the machine.

Also, the time is specified in seconds, during which the entire operation for writing data was performed. Experiment numbers correspond to the following parameters:

- Experiment 1: 1000000 records of 1000 characters each.
- Experiment 2: 100,000 records of 10,000 characters each.
- Experiment 3: 10,000 records of 100,000 characters each.
- Experiment 4: 1000 records of 1,000,000 characters each.

Tables 1-3 show the comparison of MongoDB test results with different number and size of records.

Table 01. MongoDB test results using 1 shard

Experiment #		1	2	3	4
CPU usage, %	**Shard #1**	20-30	1.3-3	1.3-3	1.3-3
	Mongos	20-30	1.3-3	1.3-3	1.3-3
RAM usage, MB	**Shard #1**	407/1497	420/1497	440/1497	451/1497
	Mongos	181/489	182/489	182/489	192/489
Writing time, s		897	803	999	994

Table 02. MongoDB test results using 2 shards

Experiment #		1	2	3	4
CPU usage, %	**Shard #1**	20-30	1.3-3	1.3-3	1.3-3
	Shard #2	20-30	1.3-3	1.3-3	1.3-3
	Mongos	20-30	1.3-3	1.3-3	1.3-3
RAM usage, MB	**Shard #1**	522/1497	480/1497	431/1497	410/1497
	Shard #2	491/1497	483/1497	426/1497	402/1497
	Mongos	182/489	182/489	182/489	196/489
Writing time, s		895	603	463	570

Table 03. MongoDB test results using 3 shards

Experiment #		1	2	3	4
CPU usage, %	**Shard #1**	20-30	1.3-3	1.3-3	1.3-3
	Shard #2	20-30	1.3-3	1.3-3	1.3-3
	Shard #3	20-30	1.3-3	1.3-3	1.3-3
	Mongos	20-30	1.3-3	1.3-3	1.3-3
RAM usage, MB	**Shard #1**	468/1497	456/1497	474/1497	417/1497
	Shard #2	443/1497	436/1497	444/1497	409/1497
	Shard #3	449/1497	391/1497	442/1497	384/1497
	Mongos	181/489	181/489	182/489	195/489
Writing time, s		843	343	316	418

Tables 4-6 show similar comparisons for Cassandra. For Cassandra, the tables also contain information about the errors received during the recording:

- Timed out errors - the number of errors of the form "ResponseError: Operation timed out - received only 0 responses."
- Tried for query failed errors - the number of errors of the form "OperationTimedOutError: The host 19 timeout 12000 ms. See innerErrors."

http://dx.doi.org/10.15405/epsbs.2017.12.26
Corresponding Author: Evgeny Nikulchev
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

Table 04. Cassandra test results using 1 shard

Experiment #		1	2	3	4
CPU usage, %	Shard #1	40-60	20-40	1-2	1-2
RAM usage, MB	Shard #1	407/1497	420/1497	440/1497	451/1497
Writing time, s		960	185	408	440
Timed out errors		4	14	116	138
Tried for query failed errors		4	2	11	10

Table 05. Cassandra test results using 2 shards

Experiment #		1	2	3	4
CPU usage, %	Shard #1	30-50	30-60	1-2	1-2
	Shard #2	10-20	30-60	1-2	1-2
RAM usage, MB	Shard #1	1421/2001	1406/2001	1398/2001	1388/2001
	Shard #2	1376/2001	1377/2001	1384/2001	1380/2001
Writing time, s		890	134	328	421
Timed out errors		6	1	96	144
Tried for query failed errors		0	0	1	0

Table 06. Cassandra test results using 3 shards

Experiment #		1	2	3	4
CPU usage, %	Shard #1	10-20	1-2	1-2	1-2
	Shard #2	20-30	1-2	1-2	1-2
	Shard #3	5-10	1-2	1-2	1-2
RAM usage, MB	Shard #1	1404/2001	1388/2001	1389/2001	1399/2001
	Shard #2	1384/2001	1376/2001	1379/2001	1362/2001
	Shard #3	1368/2001	1371/2001	1381/2001	1362/2001
Writing time, s		822	167	180	141
Timed out errors		3	3	56	50
Tried for query failed errors		0	0	0	0

With the results of testing it is visible that Cassandra demands much more RAM than MongoDB, but copes with record of files faster. The load on the processor is approximately the same for both DBMSs.

Figures 1-4 show graphs of the recording time versus the number of shards for each of the four experiments. Vertical axis is the recording time in seconds.

Figure 1 shows that recording a small amount of information (1000 characters) takes approximately the same amount of time in the case of both DBMSs. Only on three shards there is a noticeable increase in speed in Cassandra compared to MongoDB.

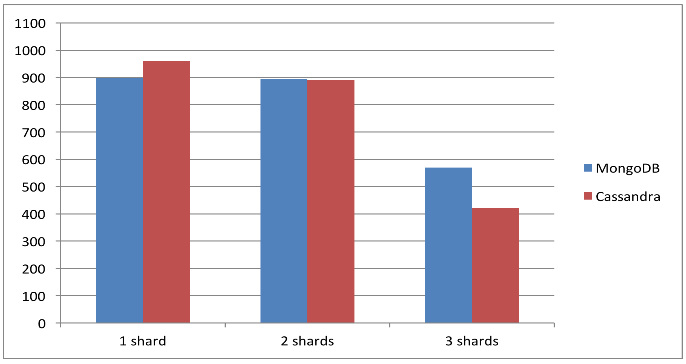

Figure 01. 1,000,000 entries of 1000 characters each

Figure 2 shows that during the experiment with the selected volume of records of 10,000 characters Cassandra does not show a speed increase. MongoDB, on the contrary, noticeably accelerates with the increase in the number of nodes.

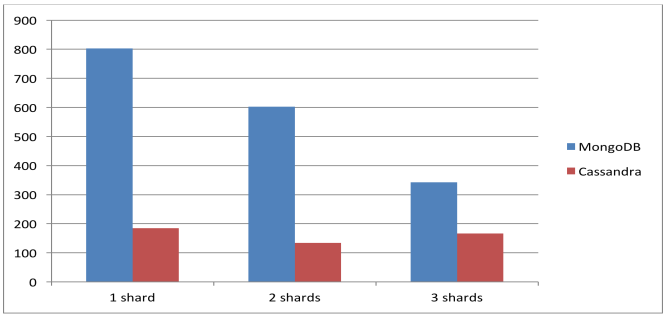

Figure 02. 100,000 entries of 10,000 characters each

Figures 3 and 4 show that during the recording of data of 100,000 and 1,000,000 characters, the increase in the recording speed with increasing number of shards becomes more pronounced.

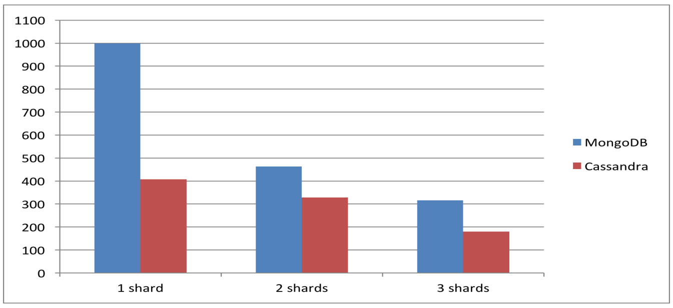

Figure 03. 10,000 records of 100,000 characters each

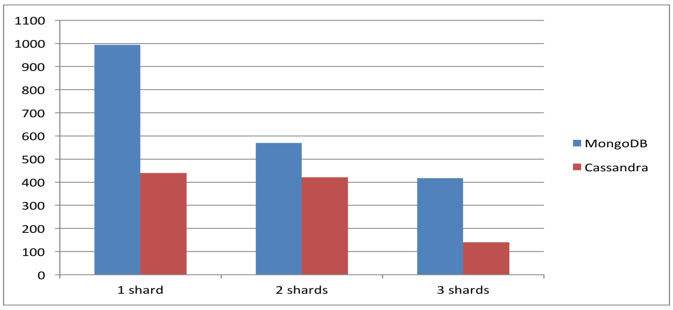

Figure 04. 1000 entries of 1,000,000 characters each

From the graphs, we can conclude that with an increase in the number of shards, MongoDB gets a greater relative speed gain than Cassandra.

The configuration of the Cassandra was much more complicated than the configuration of the MongoDB, therefore Cassandra is more demanding for the qualification of a specialist.

Despite the higher speed performance of Cassandra, the complexity of its configuration remains a major drawback in the context of the task at hand. MongoDB meets all the specified criteria, so to solve the problem of storing data of the web platform; the selection should be stopped on this DBMS.

7. Conclusion

The use of modern web technologies can be a promising tool on which the future development of psychological research depends. Therefore, for psychologists in Russia it is very important to have their own tool for conducting psychological research (Zakharov, Nikulchev, Ilin, & Ismatullina, 2017).

An experiment was conducted to assess the applicability of MongoDB and Cassandra NoSQL solutions for data storage. Both systems are scalable and meet the requirements, but due to the high qualification required to support the Cassandra DBMS, MongoDB becomes the preferred solution.

The results of the study will form the basis for the development of a web platform for psychological research. It should be noted that setting up an experiment using the Vagrant tool can increase the reproducibility of the experiments. Also, the results of the work can be useful in creating similar solutions aimed at collecting and analyzing large amounts of data.

Acknowledgments

This research is supported by the RFBR grant no. 17-29-02198.

References

Abramova, V., & Bernardino, J. (2013). NoSQL databases: MongoDB vs cassandra. roceedings of the International C* Conference on Computer Science and Software Engineering, 14-22.

Corbellini, A., Mateos, C., Zunino, A., Godoy, D., & Schiaffino, S. (2017). Persisting big-data: The NoSQL landscape. *Information Systems*, 63, 1-23.

Dede, E., Govindaraju, M., Gunter, D., Canon, R., & Ramakrishnan, L. (2013). Performance Evaluation of a MongoDB and Hadoop Platform for Scientific Data Analysis. Science Cloud '13 Proceedings of the 4th ACM workshop on Scientific cloud computing, 13-20.

Gorton, I., & Klein, J. (2014). Distribution, Data, Deployment: Software Architecture Convergence in Big Data Systems. *IEEE Software*, 32(3), 78-85.

Gudivada, V., Rao, D., & Raghavan, V. (2014). NoSQL Systems for Big Data Management. IEEE World Congress on Services 2014, 190-197.

Han, J., Haihong, E., Guan, L., & Jian, D. (2011). Survey on NoSQL database. 2011 6th International Conference on Pervasive Computing and Applications (ICPCA).

Ismatullina, V., & Voronin, I. (2017). Individual Differences in the Relationship between Temperament and Planning Ability in Adolescents. Procedia - Social and Behavioral Sciences, 237, 1455-1461.

Ismatullina, V., Zakharov, I., Nikulchev, E., & Malykh, S. (2016). Computerized tools in psychology: cross cultural and genetically informative studies of memory. *ITM Web of Conferences*, 6, 03005

Lakshman, A., & Malik, P. (2010). Cassandra - A Decentralized Structured Storage System. *ACM SIGOPS Operating Systems Review*, 44(2), 35-40.

Luciana, M. (2003). Practitioner Review: Computerized assessment of neuropsychological function in children: clinical and research applications of the Cambridge Neuropsychological Testing Automated Battery (CANTAB). *Journal of Child Psychology and Psychiatry*, 44(5), 649-663.

Nikulchev E., Pluzhnik E., Biryukov D., Lukyanchikov O. & Payain S. (2015). Experimental Study of the Cloud Architecture Selection for Effective Big Data Processing. *International Journal of Advanced Computer Science and Applications*, 6(6), 22-26.

Rimfeld, K., Shakeshaft, N., Malanchini, M., Rodic, M., Selzam, S., Schofield, K., . . . Plomin, R. (2017). Phenotypic and genetic evidence for a unifactorial structure of spatial abilities. Proceedings of the National Academy of Sciences, 114(10), 2777-2782.

Stillwell, M., & Coutinho, J. (2015). A DevOps approach to integration of software components in an EU research project. 1-6.

The number of pupils and personnel of educational institutions of the Russian Federation (The forecast up to 2020 year and score trends up to the year 2030). (2013). Tsentr sotsial'nogo prognozirovaniia i marketinga, 164.

Tikhomirova, T., & Malykh, S. (2016). The relationship of non-verbal intelligence and success in mathematics at primary school age: a longitudinal study. T*heoretical and Experimental Psychology*, 9(4), 6-22.

Verbitskaya, L., Zinchenko, Y., Malykh, S., & Tikhomirova, T. (2017). Cognitive foundations of successful teaching of the Russian language: a cross-cultural study. *Voprosy Psikhologii*, 1, 26-40.

Voronin, I., Ismatullina, V., Zaharov, I., & Malykh, S. (2016). The nature of the relationships between personality and cognitive characteristics. *SHS Web of Conferences*, 29, 02043

Zakharov, I., Ismatullina, V., & Voronin, I. (2016). Pattern recognition memory and intelligence in adolescence. *International Journal of Psychophysiology* (108), 160.

Zakharov, I., Nikulchev, E., Ilin, D., & Ismatullina, V. (2017). Web-based Platform for Psychology Research. *ITM Web of Conferences*, 10, 04006.