International Journal of Engineering & Technology, 7 (2) (2018) 902-906

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i2.12113
Research paper

A literature review on NoSQL database for big data processing

Md. Razu Ahmed [1], Mst. Arifa Khatun [1], Md. Asraf Ali [1] *, Kenneth Sundaraj [2]

[1] Department of Software Engineering, Daffodil International University, Dhaka, Bangladesh
[2] Faculty of Electronics and Computer Engineering, Universiti Teknikal Malaysia Melaka, Malaysia
**Corresponding author E-mail: asraf.swe@diu.edu.bd*

Abstract

Objective: Aim of the present study was to literature review on the NoSQL Database for Big Data processing including the structural issues and the real-time data mining techniques to extract the estimated valuable information.
Methods: We searched the Springer Link and IEEE Xplore online databases for articles published in English language during the last seven years (between January 2011 and December 2017). We specifically searched for two keywords ("NoSQL" and "Big Data") to find the articles. The inclusion criteria were articles on the use of performance comparison on valuable information processing in the field of Big Data through NoSQL databases.
Results: In the 18 selected articles, this review identified 8 articles which provided various suitable recommendations on NoSQL databases for specific area focus on the value chain of Big Data, 5 articles described the performance comparison of different NoSQL databases, 2 articles presented the background of basics characteristics data model for NoSQL, 1 article denoted the storage in respect of cloud computing and 2 articles focused the transactions of NoSQL.
Conclusion: In this literature, we presented the NoSQL databases for Big Data processing including its transactional and structural issues. Additionally, we highlight research directions and challenges in relation to Big Data processing. Therefore, we believe that the information contained in this review will incredible support and guide the progress of the Big Data processing.

Keywords: *Big Data; Data Processing; Hadoop; Mongo DB; NoSQL.*

1. Introduction

We are living an era of data ocean. The past 25 years, Data has raised in a massive scale in diverse fields including software based medical rehabilitation system [1] and sports coaching [2]. According to the report of International Data Corporation (IDC), the overall created data in the world will reach 44 ZB or trillion gigabytes during the time of 2013 to 2020 [3]. This vast amount of data refers to a Big Data that is a global buzzword. However, if we do not process Big Data, the outcome of their invisible data would be missed. It is one of the big problem for Big Data analysis and giant companies [3]. It cannot be effectively & efficiently managed using traditional database management tools [4]. To handle this problem, users have a number of options how to reach the problem related with such data. For example, to store and process vast scale datasets users can use various database technology including NoSQL databases [5]. NoSQL originally started off a simple combination of two words 'No' and 'SQL' explained as the ellipsis of not only SQL [6]. Hence, NoSQL is a generic term used to refer to any data store or process that does not follow the traditional model of relational database management system [7]. There are 4 basic types of NoSQL databases includes key-value store, document-based store, and graph-based store [8].The key-value store NoSQL database basically, uses a hash table in which there exists a unique key and value to a specific data. The values are identified and retrieved through a key, and stored values can be numbers, strings, JSON, XML, HTML, binaries, images, videos and few others [7]. Document-Store NoSQL Database, stores each record and data within a single document. A document-store NoSQL database is used for storing, recovering, and handling semi-structured data [8]. In column-oriented NoSQL database, data is stored in cells grouped in columns of data. Columns are logically grouped into column families. A Graph-based NoSQL database that uses relationships and nodes to represent and store data. NoSQL database system are emerging beside main internet and IT company, such as Google, Amazon, Facebook, Alibaba, IBM; which company are dealing with huge amount of data with traditional Relational Database System could not handle. Therefore, the aim of the study was to help users, especially to obtain an independent understanding of the strengths and weakness of NoSQL database approach and where we able to improve for managing huge volume of data.

2. Methods

2.1. Article searching procedure

We used a systematic searching procedure to identify all of the available articles that discuss the storing and managing data for Big Data situation using NoSQL databases. In our systematic searching procedure, we searched two keywords from the Springer Link and IEEE Xplore digital databases in order to assess the article. Firstly, we used the keyword "NoSQL" to find journal articles published in the English language between years 2011 to 2017. We then used the keyword "Big Data" within obtained set of search results to further narrow the set of analyzed journal article.

2.2. Article inclusion and exclusion benchmark

For the final preference of articles that applied the NoSQL database system for Big Data management. We used some benchmark to include and exclude articles from the set of articles that were selected through the search of IEEE Explore and Springer Link online databases. To include and exclude articles from the set of articles found through our systematic searching technique, we read the title, abstract, methodology and results of each article. We considered only those articles that were written in English and that used NoSQL Databases. The exclusion criteria were the following: 1) Article that applied NoSQL database other than Big Data management, 2) Article that managed Big Data using other than NoSQL database.

2.3. Data extraction

We carefully read and analyzed all of included articles to minutes the key information. We followed a standard data extraction from for the particular analysis of each article. Two of the authors of this study (RA and AK) used our designed standard data extraction form to track the key information and compared them in order to confirm the accuracy of the extracted records. Each article was evaluated for the following key information: (1) Performance comparison for different NoSQL databases for Big Data processing, (2) Overview of different NoSQL databases and Big Data processing technologies, (3) Database Engine Ranking, (4) Data Models of NoSQL Databases, (5) Transactions on NoSQL databases.

2.4. Research questions

The final set of articles was used to answer the following questions: 1) Which is the best NoSQL Databases system for Big Data Processing? 2) What is the main structural issue in the Big Data processing? 3) Are there any transactions possible on NoSQL Databases?

3. Results

3.1. Article search results

We used our systematic article searching procedure and found 18 articles that have published in reputed journals and conferences between January 2011 and December 2017. We then scanned all the articles particularly and identified the key points of each. We search for two keywords ("NoSQL Database" and "Big Data") to find of the articles. The search of the Springer Link and IEEE Xplore electronics databases using this keyword "NoSQL" retrieved 395 and 568 articles respectively.

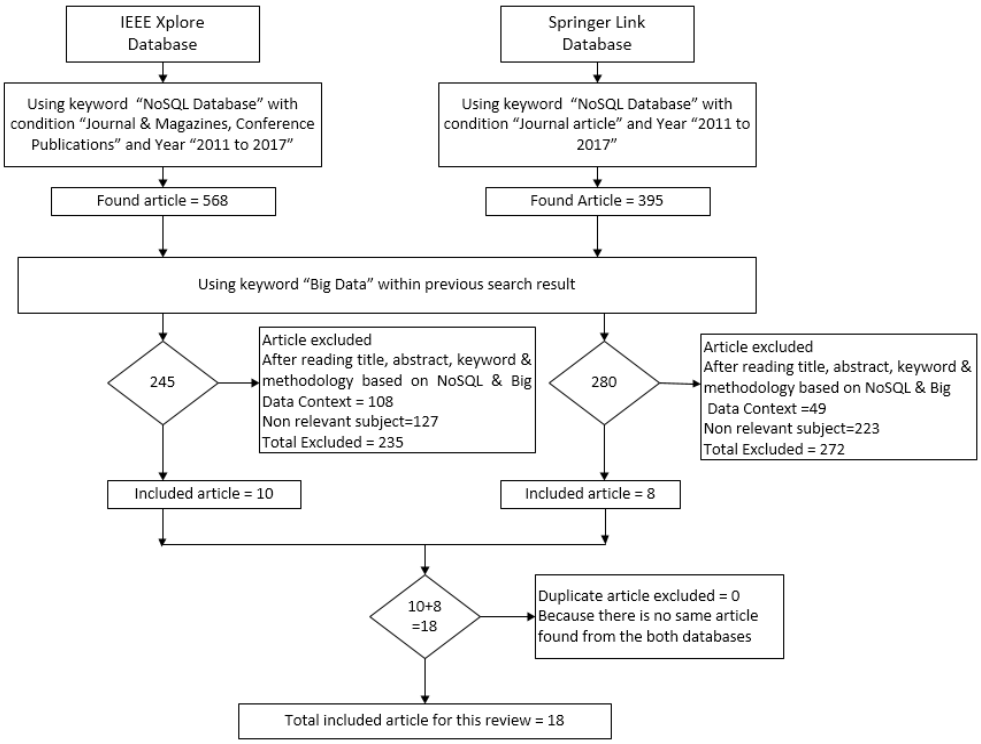

Fig. 1: Article Search Results.

We then refined search using the keyword "Big Data", and this search resulted in the retrieval of 280 and 245 from the Springer Link and IEEE Xplore databases, particularly. We then read the title, abstract, keyword, and methodology of each article, and based on our inclusion & exclusion criteria, we selected 10 articles from the IEEE Xplore database and 8 articles from the Springer Link Database for analyzing the present study. The article search result is summarized in Figure 1. Thus, as a result of this searching procedure, a total of 18 articles that discussed NoSQL database for Big Data management.

3.2. Descriptive analysis

From the collected of 18 articles, we found 8 articles [8-15] that were related to general overview of NoSQL databases and Big Data, 5 articles [16-20] that discussed performance comparison and evaluation
between different NoSQL databases, 2 articles [21-22] that were associated to NoSQL data model and classifying NoSQL databases based on the Consistency, Availability, tolerance of network Partition of database which is known as CAP theorem, [23] this research has concentrated on the storage aspect of cloud computing systems using NoSQL. Other
two [24-25] study reported the SQL query condition transformation to any NoSQL based database using Espresso Heuristic algorithm and analyzed the effects of transaction on data consistency & efficiency. However, the outcomes of the 18 articles on NoSQL used in Big Data processing are summarized in Table 1.

Table 1: Key Points of Each Article

Reference	Approach	Study
[8-15]	General overview of NoSQL and Big Data	This study provides a recommendation on the suitable databases for specific type of application requirement & focus on the value chain of Big Data.
[16-20]	Performance comparison and evaluation on different NoSQL Databases.	Discussion on NoSQL best use cases and NoSQL Databases performance measurement.
[21-22]	Classifying NoSQL Databases according to the CAP theorem and Data Model.	This study describes the background basics characteristics data model of NoSQL
[23]	DB-Engine Ranking	This research has concentrated on the storage aspect of cloud computing systems, in particular, NoSQL Databases
[24-25]	Transactional for MongoDB, Riak and NoSQL 's SQL condition based on Espresso Heuristic algorithm	To analyzed the effects of transaction on data consistency and efficiency and SQL query condition transformation for any NoSQL Databases

3.3. Research question 1: which is the best NoSQL databases system for big data processing?

Two studies [17, 18] provided a recommendation on the suitable NoSQL databases for specific type of applications. E.Tang et al. [18] experimented between five NoSQL(Redis, MongoDB, Couchbase, Cassandra, HBase) databases based on database type and popularity, and they created a 4-node cluster for each database in experiment (using YCSB = Yahoo! Cloud Serving Benchmark). They showed loading time while 100000 records load into databases. Redis got top performance which is 1.31 times quicker than MongoDB . Then, two column-based databases, Hbase and Cassandra, were 1.86 times and 1.71 times lengthier than Redis. The poor performance was presented by Couchbase database [18]. And they were consider workload execution whereas Redis also displayed best performance with the average time execution 1.1 seconds. Other study [17] described into their article, If data exists in XML format and need to achieve high level consistency then MongoDB would provide the best solution, If data are unstructured and need to high performances then Redis would be the best solution, and If processing data are a high volume of data then Cassandra would be the best solution.

3.4. Research question 2: what is the structural issue in the big data processing?

There is some structural issue of Big Data processing. Important things are how to effectively collect data and store data, and how to worth for it. These two things are definitely important for Big Data processing. Based on the above mentioned point, we found 2 studies [10, 13] that discussed the Big Data issues and challenges. M. Chen et al. [13] discussed in their survey, The world of Big Data is heterogeneous and continuously it changing alongside with IoT. Most of the data are having also wireless data, but there is no permanent gateway to collect data from this wireless data. Moreover, the study [10] noted that Big Data volume grows so large and diverse and all data does not need to store for analysis rather how it is recognized to deal with it.

3.5. Research question 3: are there any transactions possible on NoSQL databases?

Of the 18 selected studies, two studies [24, 25] focused on the ACID transactions. The primary models of NoSQL do not support ACID transaction but Gonzalez-Aparicio et al. [25] developed a new transaction system and implemented into NoSQL databases for MongoDB and Riak for ACID transaction; These transaction system allows join operation, provide high scalability and concurrency of transaction. Other study [24] described the SQL syntax converting into conditional expression of a NoSQL database.

4. Discussion

We studied particularly the selected 18 articles on "NoSQL & Big Data related technologies" for the analysis of the present processes to identify future research in the area. The current study presents a summary of the data found in the literature that focused on performance, strengths and weakness of NoSQL databases for Big Data Processing. We summarized each articles of this paper are presented as follows:

4.1. Characteristics of NoSQL databases

Most of the traditional database system are based on transactions. These transactional features are also familiar as ACID (Atomicity, Consistency, Isolation, Durability) [26]. However, Big transactional process does not work properly with ACID system [27]. Hence, ACID system shown to be a problem in different distributed systems that are not fully solvable. Therefore, Eric Brewer [28] introduced the CAP theorem (Figure 1) which is more efficient in different distributed systems. But, later the study [26] noted that the CAP theorem is accepted only two attributes among the three requirements (AP, CP, CA) for Big Data processing at a time. The more details are following:

- Available and Partition-Tolerant (AP): Achieve "eventual consistency" through reiteration and authentication. Example: Voldemort, Couch DB, Cassandra etc.
- Consistent and Partition-Tolerant (CP): CP system have trouble with availability whereas keeping data consistent across partitioned nodes. Example: MongoDB, Redis, BigTable etc.
- Consistent and Available (CA): CA system have problem with partitions and typically deal with replication. Example: Vertica, MySQL etc.

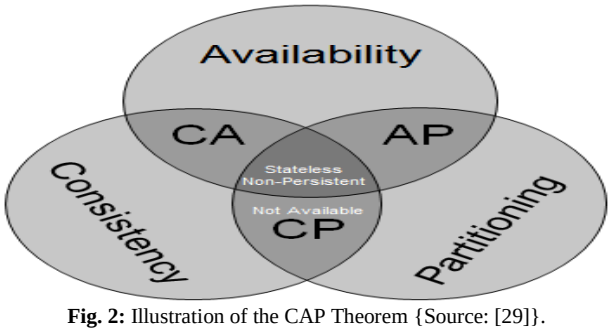

Fig. 2: Illustration of the CAP Theorem {Source: [29]}.

By the above discussion, it is very important to know about The CAP theorem for designing any distributed system. For example, when transactional and ACID issues are coming in NoSQL database, there is no other option without CAP theorem. Later, Gonzalez-Aparicio et al [25] developed a new transaction system using three components: i) Transmission Processing Engine (TPE), ii) Data Management Store (DMS), and iii) Times Stamp Manager (TSM). Also, they implemented their developed technique into NoSQL databases for MongoDB and Riak; where TPE allows join operation which are not previously supported in NoSQL, whereas DMS and TSM are applied in order to provide high scalability and concurrency of transactions. Other study [24] described the Espresso heuristic algorithm for converting SQL syntax to a conditional expression of a specific NoSQL Database (MongoDB). However, some of NoSQL databases support ACID transaction, such as

Mark Logic [30] is a solution that works like relational and NoSQL databases.

4.2. Comparison of NoSQL databases

In this present study, we provide evaluation of four categories NoSQL databases including document based (MongoDB, CouchBase), column based (Cassandra, HBase), key value base (Redis) and graph based (Neoj4). Moreover, MongoDB, CouchBase, Cassandra, HBase, Redis are the most prominent NoSQL databases that are evaluated using YCSB [16, 18]. Surya et al. [16] showed their experiment for MongoDB, CouchBase, Cassandra, HBase using YCSB, where Cassandra had a better performance than others while the execution of workload is 50% read and 50% write. They also found that HBase had a better performance with a small datasets of database, whereas MongoDB had the best performance while working with 100% read and 100% blind write [16]. EnqingTang et al. [18] described their experiment for performance measurement between different NoSQL databases including Radis, MongoDB, CouchBase using YCSB and noted that Redis is particularly appropriate for loading and executing workloads for small datasets, but Redis have poor performance for the issues of vast amount of datasets. The study [8] worked with various NoSQL databases including Redis, MongoDB, CouchBase and noted that Redis is the best suited for the analysis of small amount of datasets in order to get high performance, and for processing large amount of data MongoDB is the best choice, whereas CouchBase is better for fault-tolerant database environment. However, CoucheBase is applied for working concurrently read and write operation in large scale data sets but performance is not good enough, in this case it is recommended to apply Cassandra [17].

4.3. Big data processing using NoSQL

Generally, Big Data is an issue when the size of the dataset crosses the ability of existing software in order to analyze dataset including data collecting, processing, retrieving and managing. According to the study of [31] the Big Data frequently described using 5 Vs(Volume, Velocity, Variety, Veracity, Value) are following:

- Volume (great volume): The large amount of data sets are created in every second. As a result, data sets are used to increase time over time. This increasingly makes data sets too large in order to store and analyses for traditional database technology. But, considering the Big Data technology, we are able to store and use these kind of large data sets with the help of distributed systems, where parts of the data is stored in different locations and brought together by software.
- Velocity (rapid procreation): Velocity refers to the speed at which new data is generated and the speed at which data moves around. Big data technology allows us to analyze the data while it is being generated, even without putting it into databases.
- Variety (various types): With the Big Data technology, we are able to join differed types of data (structured and unstructured) including messages, social media conversations, photos, sensor data, video or voice recordings and bring them together with more traditional, structured data.
- Veracity: It refers to the messiness or constancy of the data. With many forms of big data, quality and accuracy are less controllable (just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content) but Big Data and analytics technology allows us to work with these type of data. The volumes often make up for the lack of quality or accuracy.
- Value (huge value but very low density): Value is also important to take into account when looking at Big Data. It is well and good having access to Big Data but unless turns it into a value otherwise it is useless. It is noted that businesses make a business case for any attempt to collect and leverage big data. It is so easy to fall into the buzz trap and embark on Big Data initiatives without a clear understanding of costs and benefits.

  Moreover, the study [32] added another 2 Vs (Variability, Visualization) with the above noted 5V's are following:
- Variability: Variability is different from variety. For example, a coffee shop may offer 6 different blends of coffee, but if we get the same blend every day and it tastes different every day, that is variability. The same is true for data sets, if the meaning is constantly changing, it can have a huge impact on data homogenization.
- Visualization: Visualization is complex in today's world. Using charts and graphs to visualize huge amounts of complex data is much more effective in conveying meaning than spreadsheets and reports chock-full of numbers and formulas.

However, the amount of data has been generating more and more and therefore data sets analysis become more complex. This challenge is not only collecting and managing vast amount of data, it also more challenging to extract valuable data [33]. Hence, extracting valuable data is important issue that need to process by four phases are presented in Figure 4.

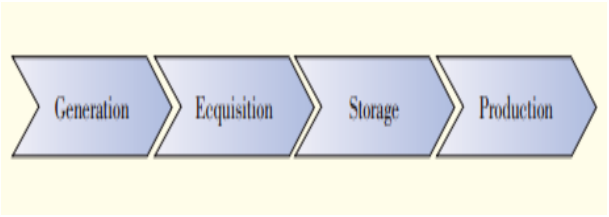

Fig. 3: Big Data Life Cycle.

In the generation phase of Big Data processing, various type of raw data sets are created by different sources. The second phase is acquisition that involve for data collecting and pre-processing in order to transmit the data sets as the Big Data to the storage phase. Generally, storage phase is used to store the Big Data using Distributed file system (DFS). Recently, there are different types of DFS available for big data processing such as HDFS, GFS, and TFS. Finally, the Big Data production is the most important for analysis of Big Data sets which is used to perform through batch-based (Map-Reduce), BSP-based, or stream based data processing techniques [34]. However, real-time analysis and small data sets analysis are complex in Hadoop Distributed System. Thus, the real time and non-files data set analysis is used to perform through NoSQL database. Hence, to handle the vast amount of data sets, it is require to consider some issues and challenges are following:

- Extremely difficult to effectively collecting and processing of big data.
- If data volume grows so large and diverse, then there is no particular technique to deal with it.
- Slower Decision making, automatic data mining process will be more improved.
- To collecting wireless data, there is no stable gateway.

In summary, we highlight research directions and challenges in relation to Big Data processing and storage management system by NoSQL. Which is emerging impact of Big Data analysis. In NoSQL databases, transactional issue into NoSQL database and structural gap between cloud infrastructures can be more improved. Here, we describe major databases features that require further research in terms of Big Data management.

5. Conclusion

In this paper, we have described NoSQL database in Big Data technologies. We categorized different NoSQL databases, particularly CAP theorem and then we discussed about the Big Data life cycle. Moreover, we have discussed the strengths, weakness, comparison and evaluation on various NoSQL databases. The pointing out from our discussion would be helpful to the business leader for selecting an appropriate NoSQL database in order to store and manage the

Big Data. In addition, we have focused research directions and challenges in relation to the Big Data in storage and management system through NoSQL. Although, several techniques have been applied for Big Data processing through NoSQL, their use for real-time data mining process and to extract estimated valuable information may still be compromised by the factor highlighted in the present study.

6. Competing interest

The authors declared that they have no competitive interest

References

[1] N. U. Ahamed, K. Sundaraj, R. B. Ahmad, M. Rahman, and A. Ali, A Framework for the Development of Measurement and Quality Assurance in Software-Based Medical Rehabilitation Systems, Procedia Engineering. 2012; 41: 53–60. https://doi.org/10.1016/j.proeng.2012.07.142.

[2] N. M. Khan, Realtime coaching system, U. S Patent No. 8,279,051, 2012.

[3] A. A. Safaei, Real-time processing of streaming big data, Real-Time System. 2017 53(1):1–44. https://doi.org/10.1007/s11241-016-9257-0.

[4] InFocus Blog | Dell EMC Services. Big Data and NoSQL: The Problem with Relational Databases. https://infocus.emc.com/april_reeve/big-data-and-nosql-the-problem-with-relational-databases/. Accessed August 7, 2017.

[5] J. Pokorny, New Database Architectures: Steps towards Big Data Processing, Proc. IADIS Eur. Conf. Data Min. (ECDM'13), António Palma dos Reis Ajith P. Abraham Eds., IADIS Press. 2013 3–10.

[6] M. Stonebraker, SQL databases v. NoSQL databases, Commun. ACM. 2010 53(4):10-11. https://doi.org/10.1145/1721654.1721659.

[7] Tech Republic. 10 things you should know about NoSQL databases. http://www.techrepublic.com/blog/10-things/10-things-you-should-know-about-nosql-databases/. Accessed August 8, 2017.

[8] J. Bhogal and I. Choksi, Handling Big Data Using NoSQL. in Proceedings - IEEE 29th International Conference on Advanced Information Networking and Applications Workshops, WAINA 2015, 393–398: IEEE

[9] R. Zafar, E. Yafi, M. F. Zuhairi, and H. Dao, Big Data: The NoSQL and RDBMS review, in ICICTM 2016 - Proceedings of the first International Conference on Information and Communication Technology, 2017; 120–126.

[10] I. Chebbi, W. Boulila, and I. R. Farah, Big Data: Concepts, Challenges and Applications, Springer, Cham, 2015 638–647.

[11] S. D. Kuznetsov and a. V. Poskonin, NoSQL data management systems. Program. Comput. Softw. 2014 40(6): 323–332. https://doi.org/10.1134/S0361768814060152.

[12] V. Rajaraman, Big data analytics, Resonance. 2016 21(8):695–716.

[13] M. Chen, S. Mao, and Y. Liu, Big data: A survey, in Mobile Networks and Applications, 2014; 19 (2):171-209. https://doi.org/10.1007/s11036-013-0489-0.

[14] F. Gessert, W. Wingerath, S. Friedrich, and N. Ritter, NoSQL database systems: a survey and decision guidance, Comput. Sci. - Res. Dev. 2017; 32(3–4): 353–365.

[15] L. Wu, L. Yuan, and J. You, Survey of Large-Scale Data Management Systems for Big Data Applications, J. Comput. Sci. Technol., 2015; 30(1): 163–183. https://doi.org/10.1007/s11390-015-1511-8.

[16] S. Swaminathan and R. Elmasri, Quantitative analysis of scalable nosql databases, 2016 IEEE International Congress on Big Data (BigData Congress), 2016: IEEE.

[17] P. Srivastava, S. Goyal, and A. Kumar, Analysis of various NoSql database, 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), 2015: IEEE. https://doi.org/10.1109/ICGCIoT.2015.7380523.

[18] E. Tang and Y. Fan, Performance Comparison between Five NoSQL Databases, in 2016 seventh International Conference on Cloud Computing and Big Data (CCBD), 2016; 105–109: IEEE.

[19] D. Seybold, N. Wagner, B. Erb, and J. Domaschka, Is elasticity of scalable databases a Myth? , in 2016 IEEE International Conference on Big Data (Big Data), 2016 2827–2836: IEEE.

[20] V. N. Gudivada, D. Rao, and V. V. Raghavan, NoSQL Systems for Big Data Management, 2014 IEEE World Congr. Serv., 2014; 190–197: IEEE. https://doi.org/10.1109/SERVICES.2014.42.

[21] A. Mohan, M. Ebrahimi, S. Lu, and A. Kotov, A NoSQL Data Model for Scalable Big Data Workflow Execution, in 2016 IEEE International Congress on Big Data (BigData Congress), 2016;52–59:IEEE.

[22] J. Han, E. Haihong, G. Le, and J. Du, Survey on NoSQL database, in Proceedings - 2011 6th International Conference on Pervasive Computing and Applications, ICPCA 2011, 2011; 363–366:IEEE.

[23] K. Grolinger, W. a Higashino, A. Tiwari, and M. A. Capretz, Data management in cloud environments: NoSQL and NewSQL data stores, J. Cloud Comput. Adv. Syst. Appl., 2013 2, 22. https://doi.org/10.1186/2192-113X-2-22.

[24] Changqing Li and Jianhua Gu, A distinctive transformation approach of NoSQL's SQL conditions based on Espresso, in 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2017; 61–69: IEEE.

[25] M. T. Gonzalez-Aparicio, A. Ogunyadeka, M. Younas, J. Tuya, and R. Casado,Transaction processing in consistency-aware user's applications deployed on NoSQL databases, Human-centric Comput. Inf. Sci., 2017; 7(1): seven.

[26] j. Han, E. Haihong, G. Le, and J. Du, Survey on NoSQL database, in Proceedings - 2011 6th International Conference on Pervasive Computing and Applications, ICPCA 2011, 2011;363–366:IEEE.

[27] A. B. M. Moniruzzaman and S. A. Hossain, NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison, 2013; arXiv preprint arXiv: 1307.0191.

[28] E.A. Brewer. Towards robust distributed systems. InPODC, 2000; seven.

[29] TheTechSolo. CAP Theorem. https://thetechsolo.files.wordpress.com/2016/02/captheorem.png?w=640. Accessed September 19, 2017.

[30] Integrate Data Silos with a NoSQL Database | What Is MarkLogic. http://www.marklogic.com/what-is-marklogic/.Accessed:August 8, 2017.

[31] LinkedIn Pulse. B. Marr, Big Data: The five Vs Everyone Must Know, 2014.

[32] Affect Radius.The 7 V's of Big Data https://www.impactradius.com/blog/7-vs-big-data/.Accessed: September 19, 2017.

[33] S. Sagiroglu and D. Sinanc, "Big data: A review, in 2013 International Conference on Collaboration Technologies and Systems (CTS), 2013; 42–47: IEEE. https://doi.org/10.1109/CTS.2013.6567202.

[34] Distributed File System. https://www.slideshare.net/Lycca/distributed-file-system.Accessed January 3, 2018.