SOUTHEAST ASIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Volume 6, Issue 2, (Special Issues) 2021
P-ISSN: 2672-2984
E-ISSN: 2672-2992
www.sajst.org

Empirical Research on the Challenges of Distributed Databases: A Literature Review

Patrick V. Mole, and Elmar B. Noche
Pangasinan State University

Abstract - This paper presents an empirical study on the challenges of the distributed database system. The paper reflects on the understanding of what are the different issues and challenges in the past years. The literature review demonstrates the identified solutions to the specific problems, as well as the lapses of the distributed database system, and examines the state of current research on the topic and points out gaps in the existing literature. The findings of the study revealed that a distributed database system encountered a lot of issues in the past years, but this technology has been overcome most of the challenges. However, there are still issues that still persist in recent years and have yet to be addressed adequately which affect the performance of this technology. These problems provide insight for further studies and keep the future researchers as well as distributed database system implementers quite busy for some time to come.

Keywords - Distributed Database, Database, Distributed Systems

1. Introduction

In the past decade, distributed database technology has increasingly replaced centralized databases in different domains such as commerce, banking, and even in an institution [10]. A distributed database is very useful now due to promising features to manage a database. It provides a ‘‘seamless’’ interface to data that is stored on multiple computer systems [20]. According to Ceri [3], the technical features that motivate the success and growth of distributed databases have not to deal with their reduced cost, but rather with their increased benefits.

Distributed database research issues have been topics of intense study, culminating in the release of several “first generation” commercial products. Distributed database technology is expected to impact data processing the same way that centralized systems did a decade ago. Stonebraker suggests that within the next ten years, centralized database managers will be an “antique curiosity” and most organizations will move toward distributed database managers [20]. According to Pupezescu et. al (2019), some of the organizations are motivated to implement

efficient Distributed Database Systems (DDBS) or Decentralized Database Systems in their administrative environments for scalability [14]. Furthermore, several studies [20][3][13] showed that the distributed database is the new direction and the present trends of most of the organizations. Despite being popular, distributed databases are still facing problems and issues regarding the performance of this technology.

There has also been a mountain of research discussing the common problems in the distributed database [17][2][19][4]. However, the issue is not only that the commercial systems must catch up and implement the research results, but that there are still significant research problems that remain to be solved. Even the much-studied topics in past years, there is a dearth of understanding of the impact of the remaining issues and challenges to the performance of the distributed database.

This paper aims to carry out a systematic literature review of empirical evidence about the different challenges of a distributed database system. This review will focus on answering the following questions: 1) What are the different issues and challenges in implementing a distributed database system? 2) What are the solutions presented to address the issues and challenges? In retrospect, what are the challenges that still have no solution? and, 3) What are the factors of these remaining challenges affecting the performance of the distributed database system?

This paper will contribute to the understanding of the issues which affect the performance of the distributed database by reviewing the existing body of empirical research on the topic.

2. METHODOLOGY

2.1 Literature search

The findings of relevant studies in this literature review were published studies on different electronic databases such as Google Scholar, Semantic Scholar, Springer, Science Direct, ACM (Association for Computing Machinery), Research Gate, and IEEE (Institute of Electrical and Electronics Engineers). After the initial searches in electronic databases, the number of criteria is specified to select relevant studies for inclusion in the review. The paper selection process included 1) empirical study evidence regarding distributed databases and its performance, 2) published as articles journals, 3) include an abstract, and 4) the research method is demonstrated clearly. A citation was excluded if 1) it was in academic settings such as book review 2) it was a book or chapter, or 3) no empirical data was reported. The information was systematically collected including the authors, year of publication, objectives of the study, the settings, the method used (research design, data collection, and analysis methods), and the main results from each of the studies.

2.2 Identifying the Challenges in Distributed Database Systems

There are two phases in conducting the process of determining the different challenges of the distributed database. In the first phase, the journals are examined against the inclusion and exclusion criteria. The journals without empirical evidence related to the issues and challenges of the distributed database systems are excluded in the review and the remaining journals with potential are acquired to be analyzed further in the next phase. In the second phase, the challenges from the reviewed journals are listed and the overview of each issue is summarized.

2.3 Identifying the Solutions for the Challenges in Distributed Database Systems

Further investigation of the listed challenges in a distributed database system was made. The researchers further analyzed the solutions in every issue and challenges in distributed database system. A number of solutions have been identified for the challenges; however, these solutions did not resolve all the issues in the distributed database. In this phase, all of the identified challenges and their corresponding solutions are separated into two parts; 1) challenges which already have a solution, and 2) remained challenges that await solution. The additional searches through references are also conducted in this phase where the researchers further investigated the references of the initially found papers and the references made to those papers. The supplementary investigation is used to validate if the distributed database system still has no solution to the identified remaining challenges in recent years.

2.4 Identifying the Factors affecting Distributed Database Systems Performance

The eligible peer-reviewed research papers related to the remaining challenges in the distributed database were left for the review. In this phase, the papers were analyzed to determine the factors affecting the performance of the distributed database system concerning the issues that still remained.

3. RESULTS AND DISCUSSION

3.1 General Database Search

Table 1 shows the number of papers in each of the databases that were identified using the key search. The second column in table 1 contains all the results including the non-scientific writings such as magazine articles. In all, there are 67 journals examined against the inclusion and exclusion criteria. Based on the reading of abstracts, 30 journals are excluded in the review and the remaining 37 journals with potential are acquired. The remaining full texts journal articles are evaluated to check if it has empirical evidence. Finally, a total of 15 articles are excluded and 22 qualified peer-reviewed research papers on the distributed database were left for the review.

Table 1. Results from searched databases

Database	**Total number of results**	**Peer-reviewed papers**
Semantic Scholar	1027	5
IEEE	675	3
Springer	118	7
Research Gate	48	2
ACM	7	2
Science Direct	11	3
Google Scholar	4261	N/A

3.2 Issues and challenges in DDBS

Most of the reviewed papers reported common problems of distributed databases in their study and through analyzing, a total of 11 issues and challenges are identified that distributed databases encountered up to recent years. However, new issues arise with the changing technology, expanding application areas, and the experience that has been gained with the limited application of distributed database technology. Furthermore, while these problems are important ones to address, many others remain unsolved.

Table 2 displays the summaries of the reported findings related to the challenges or problems in the distributed database system

Table 2. Challenges in DDBS

Challenges	**Paper**
Distributed Concurrency Control	[15] [6] [1]
Replication Control	[15] [6] [14]
Distribution Design	[12] [6]
Deadlock Handling	[15] [16]
OS Environment	[12] [15] [6]
Distributed Query Processing	[12] [13]
Security and Privacy	[15] [6] [8] [5]
Distributed Multidatabase Systems	[12] [7]

3.3 Solutions for the Challenges in Distributed Database Systems

There have been a lot of studies concerning the different problems and challenges in a distributed database system. Table 3 shows the identified solutions in the common challenges of DDBS.

Table 3. Solutions for the challenges in DDBS

Solutions	**Paper**
Locking Techniques	[1] [15]
Fragmentation	[22] [12]
Deadlock Detection Agents	[15] [11]
Summary-schema Model	[7]

Abbas et. al highlight the locking as a technique of synchronizing the access through concurrent transactions towards database items that resolve the distributed concurrency control issue [1]. Through fragmentation, a lot of issues in DDBS can be resolved. Fragmentation is a process by which large schema records are divided into independent pieces or parts to increase the speed of distributed query processing [22]. This also enables us to achieve the distribution design and security issues in DDBS [12]. For deadlock handling issues, Krivokapic et. al presented the deadlock detection agents (DDAs). Deadlock resolution strategies determine which transaction(s) are to be aborted in order to resolve the deadlock [11]. Bright and Hurson introduce the summary-schema model which is one of the possible solutions to the problems of powerful user interfaces, effective distribution of processing, and increased semantic content in distributed multidatabase systems [7].

Almost all of the identified challenges in DDBS have already been solved, however, there are issues and challenges still persist in this recent year such as the integration of distributed databases with distributed operating systems, and the data replication control. The distributed DBMSs require modifications in how the distributed OS performs their traditional functions (e.g., task scheduling, naming, buffer management). In this context, efforts that include too much of the database functionality inside the operating system kernel or those that modify tightly-closed operating systems are likely to prove unsuccessful [12][15][6]. Besides, some of the operating systems are not supported by distributed databases. On the other side, replication of data is the method of sharing information to make sure data consistency between redundant resources, such as software or hardware components, to improve reliability, fault-tolerance, or accessibility [23]. However, many issues affect the design of a replicated database system to maintain its requirements [15][6][14]. These are the remaining challenges in a distributed database system that persist in the recent years, furthermore, new issues arise with the rapid changing of technology that may cause a distributed database system to unsolved these issues.

3.4 Factors Affecting Distributed Database Systems Performance

Based on the previous section, the unsolved challenges are the Replication Control and OS Environment. Therefore, further investigation of these unsolved issues is established. Table 3 shows the factors of the unsolved issues or the remaining challenges which affect the distributed database system performance.

Table 3. Factors affecting DDBS performance

Factors	**Paper**
Data Consistency & Reliability	[23] [22] [21]
Operating System types and version	[12] [15] [6] [21]

Many issues affecting the performance of the distributed database system in maintaining its requirements. Data Consistency and Reliability and Operating System types and the different versions based on the remaining challenges are the main issues that are considered in this paper.

3.4.1 Data Consistency and Reliability

The data consistency and reliability are the main challenges in data replication control. These challenges remain unsolved due to the changes in the common data-items from mobile users in a concurrent environment that took place simultaneously which raise a lost-update problem issue. It is a situation in which only one user change is reflected on the common data-item that result in the replication process to be incomplete or useless because it allows propagating changes of one of the users in the database. Additionally, when changes take place on a few data-items at any activity center, the propagation of changes of whole data blocks to the other sites will start which can cause a delay in data propagation. In this case, the consistency and the reliability of the data during replication are affected that resulting in the poor performance of the distributed database system [23].

3.4.2 Operating System types and version

With the rapid changing of technology, some applications are required to update to improve the stability of the software and remove outdated features. In this case, upgrading the operating system to the latest version is important to meet the changing technological needs. However, there is a negative effect in upgrading the version of the operating system to the performance of the distributed database system. According to Date's rule [15] in his twelve directives plus a fundamental principle to describe DDBS, the distributed database system can run on any kind of operating system (e.g. Windows, Linux, macOS, Android). However, (Özsu M. et al.) stated that the efforts that include too much of the database functionality inside the operating system kernel or those that modify tightly closed operating systems are likely to prove unsuccessful. Furthermore, several types of operating systems may also affect the performance of the distributed database system, especially in the mobile environment. According to Swaroop, V., and Shanker, U., the problems become more complicated in the Mobile Distributed Real-Time Database System (MDRTDS), where a database is partitioned into several smaller distributed databases as well as local database spreading at a different location. The database technology allows users to use handheld devices to link to their corporate networks, download data, work offline, and then connect to the network again to synchronize with the corporate database [21].

The summary of the analysis shows the current solutions for the challenges of the distributed database and the remaining issues. By presenting the factors of the remaining challenges affecting distributed database performance, it will now uncover the factors why these challenges still persisted in recent years.

4. Conclusion

In this paper, the discussion of the state-of-the-art in distributed database research is conducted. Specifically, the researchers (a) reviewed the different issues and challenges in implementing a distributed database system; (b) discussed the identified solutions for the specific issue and addressed the remaining challenges, and (c) addressed the factors affecting the performance of DDBS.

In this paper, the researchers figures out the common problem in implementing the distributed database system, namely distributed concurrency and replication control, distributed query processing and multidatabase system, issues in implementing to the various versions of operating systems, distribution design that lead to the security and privacy concern, and issue in detecting deadlock. In addition, these problems can be further aggravated by the changing nature of the technology on which distributed DBMSs are implemented.

The majority of the challenges in distributed databases have already been solved. Some of the identified solutions resolve numerous challenges in DDBS because most of the challenges are relevant and somewhat connected to each other.

For instance, fragmentation resolves a lot of common issues such as distributed query processing, distribution design, and security issues. As well as the summary-schema model which settles the distribution of processing and distributed multidatabase systems issues. This suggests that the distributed database is very flexible in resolving various challenges that make it even popular compared to the traditional database management. In some cases, the provided solution resolves only a particular issue. In the context of distributed concurrency control, locking technique insights to resolve this issue. Besides, the deadlock handling can be controlled through the use of deadlock detection agents. However, the replication control and the integration of distributed databases in various operating systems are the challenges that still persist in recent years. The analysis suggests that these issues are continuously developing due to the rapid changes of technology so that there is no absolute solution for these challenges.

Another important problem considered in this study was the factors of the remaining issues that affect distributed database performance. The two most common unsolved issues are the replication control and the OS environment. The factors affecting the issue in implementing a distributed database in the operating system was the different version and types of OS environment. Although distributed database systems have no problem and executing well in some operating system, it is still considered as remaining challenges in DDBS due to the different specification of OS especially in mobile environment such as android and iOS that make the distributed database to unfulfilled the lapses in this remaining issue. Furthermore, the OS is consistently providing an update to the devices. The other core process of distributed databases is data replication. It is the technique used in distributed databases to store multiple copies of a data table at different sites. The replication control arises a lot of issues such as the lost-update problem that directly affect the consistency and the reliability of data. It is now concluded that having multiple copies in multiple sites can cause a serious problem in maintaining data consistency, particularly during update operations which lead to the poor performance of distributed database systems.

REFERENCES

[1] Abbas, Q., Shafiq, H., Ahmad, I. and Tharanidharan, S., 2016, January. Concurrency control in distributed database system. In *2016 International conference on computer communication and informatics (ICCCI)* (pp. 1-4). IEEE.

[2] Bernstein, P.A. and Goodman, N., 1981. Concurrency control in distributed database systems. *ACM Computing Surveys (CSUR)*, *13*(2), pp.185-221.

[3] Ceri, S., 1988. Directions in distributed databases. In *GI—18. Jahrestagung II* (pp. 633-638). Springer, Berlin, Heidelberg.

[4] Davidson, S.B., Garcia-Molina, H. and Skeen, D., 1985. Consistency in a partitioned network: a survey. *ACM Computing Surveys (CSUR)*, *17*(3), pp.341-370.

[5] Dewitt, D. and Gray, J., 1992. Parallel database systems: the future of high performance database systems. *Communications of the ACM*, *35*(6), pp.85-98.

[6] Hiremath, D.S. and Kishor, S.B., 2016. Distributed database problem areas and approaches. *IOSR Journal of Computer Engineering (IOSR-JCE)*, *2*, pp.15-18.

[7] Hurson, A.R. and Bright, M.W., 1991. Multidatabase systems: An advanced concept in handling distributed data. In *Advances in computers* (Vol. 32, pp. 149-200). Elsevier.

[8] Karr, A.F., Fulp, W.J., Vera, F., Young, S.S., Lin, X. and Reiter, J.P., 2007. Secure,

privacy-preserving analysis of distributed databases. *Technometrics*, *49*(3), pp.335-345.

[9] Karr, A.F., Lin, X., Sanil, A.P. and Reiter, J.P., 2005. Secure regression on distributed databases. *Journal of Computational and Graphical Statistics*, *14*(2), pp.263-279.

[10] Kituta, K., Kant, S. and Agarwal, R., 2019. A systematic review on distributed databases systems and their techniques. *Journal of Theoretical and Applied Information Technology*, *96*(1), pp.236-266.

[11] Krivokapić, N., Kemper, A. and Gudes, E., 1999. Deadlock detection in distributed database systems: a new algorithm and a comparative performance analysis. *The VLDB Journal*, *8*(2), pp.79-100.

[12] Ozsu, M.T. and Valduriez, P., 1991. Distributed database systems: where are we now? *Computer*, *24*(8), pp.68-78.

[13] Pukdesree, S., Lacharoj, V. and Sirisang, P., 2010, October. An empirical study of distributed database on PC cluster computers. In *Proc. 10th WSEAS Intl. Conf. on Appl. Comp. Sc. ACS* (Vol. 10, pp. 111-115).

[14] Pupezescu, V. and Rădescu, R., 2016, June. The influence of data replication in the knowledge discovery in distributed databases process. In *2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* (pp. 1-6). IEEE.

[15] Rana, M.S., Sohel, M.K. and Arman, M.S., 2018. Distributed Database Problems Approaches and Solutions-A Study. In *International Journal of Machine Learning and Computing (IJMLC)*.

[16] Reddy, P.K. and Bhalla, S., 1993. Deadlock prevention in a distributed database system. *ACM Sigmod Record*, *22*(3), pp.40-46.

[17] Selinger, P.G. and Adiba, M., 1980, July. Access path selection in distributed data base management systems. In *Proceedings International Conference on Databases* (pp. 204-215).

[18] Shanker, U., Misra, M. and Sarje, A.K., 2008. Distributed real time database systems: background and literature review. *Distributed and parallel databases*, *23*(2), pp.127-149.

[19] Skeen, D., 1981, April. Nonblocking commit protocols. In *Proceedings of the 1981 ACM SIGMOD international conference on Management of data* (pp. 133-142).

[20] Stonebraker, M., 1989. Future trends in database systems. *IEEE Transactions on Knowledge & Data Engineering*, (1), pp.33-44.

[21] Swaroop, V. and Shanker, U., 2010, September. Mobile distributed real time database systems: A research challenges. In *2010 International Conference on Computer and Communication Technology (ICCCT)* (pp. 421-424). IEEE.

[22] Tarun, S., 2019. Distributed Database Systems Design Challenges and Countermeasures – A Study. *Journal of the Gujarat Research Society*, *21*(6), pp.875-886.

[23] Tiwari, S.K., Sharma, A.K. and Swaroop, V., 2011. Issues in Replicated data for Distributed Real-Time Database Systems. *IJCSIT) International Journal of Computer Science and Information Technologies*, *2*(4), pp.1364-1371.