Universiti Malaysia PAHANG
Engineering • Technology • Creativity

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Doctor of Philosophy in Computer Science

(Supervisor's Signature)

Full Name : ASSOCIATE PROFESSOR DR NORAZIAH AHMAD

Position : SENIOR LECTURER

Date : 25 JULY 2018

Universiti Malaysia PAHANG
Engineering • Technology • Creativity

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

(Student's Signature)

Full Name : AINUL AZILA BINTI CHE FAUZI

ID Number : PCC13001

Date : 25 JULY 2018

BINARY VOTE ASSIGNMENT ON GRID QUORUM REPLICATION TECHNIQUE WITH ASSOCIATION RULE

AINUL AZILA CHE FAUZI

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy

Faculty of Computer Systems & Software Engineering

UNIVERSITI MALAYSIA PAHANG

AUGUST 2018

No. Perolehan
126615
Tarikh
16

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisor Assoc. Prof. Dr. Noraziah Ahmad for the continuous support of my study and research, for her patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my study.

I express my deepest thanks to my family, especially to my father, Che Fauzi Bin Che Mohamood; my mother, Azizah Binti Yusoff; my sisters, Ainul Liyana and Ainul Auni and my brother, Ahmad Shahril Danial; for their patience and moral support throughout my life. Special thanks to my little angels, Aryan and Aryana for always brighten up my days and give me strength when I was feeling down. I couldn't do this without all of you.

I also want to thank my fellow mates in Faculty of Computer System and Software Engineering UMP for the stimulating discussions. Last but not least, I want to thank all best friends in UMP for the sleepless nights we were working together and for all the fun we have had in these few years.

ABSTRAK

Salah satu cabaran terbesar yang perlu dihadapi oleh pengguna pada hari ini adalah dalam penambah baikan dalam sistem pengurusan data. Organisasi perlu menyediakan data terkini kepada pengguna yang mungkin berada pada jarak yang jauh secara geografi dan untuk mengendalikan jumlah permintaan data dibahagikan di beberapa laman dalam persekitaran teragih. Oleh itu, penyimpanan, kesediaan, dan konsistensi data adalah isu-isu penting yang perlu diberi perhatian untuk membolehkan akses data yang cekap dan selamat dari laman web yang berbeza. Salah satu cara untuk menangani cabaran ini dengan berkesan adalah dengan menggunakan teknik replikasi. Replikasi adalah teknik yang berguna untuk sistem pangkalan data teragih. Melalui teknik ini, data boleh diakses dari pelbagai lokasi. Oleh itu, ia meningkatkan ketersediaan data dan akses kepada pengguna. Apabila satu pelayan gagal untuk melakukan transaksi, pengguna masih boleh mengakses data yang sama di pelayan lain. Teknik-teknik seperti Read-One-Write-All (ROWA), Hierarchical Replication Scheme (HRS) dan Branch Replication Scheme (BRS) adalah teknik popular yang digunakan untuk replikasi dan pengurusan data. Walau bagaimanapun, teknik ini mempunyai kelemahan dari segi kos komunikasi, iaitujumlah pelayan replikasi yang diperlukan untuk menyalin data. Selain itu, teknik-teknik ini juga tidak mempertimbangkan perkaitan antara data semasa proses pembahagian data. Pengetahuan mengenai perkaitan antara data dapat diekstrak dari data yang lepas dengan menggunakan teknik-teknik dalam bidang perlombongan data. Tanpa strategi yang tepat, replikasi boleh menyebabkan masa untuk melengkapkan satu transaksi menjadi tinggi. Dalam penyelidikan ini, Binary Vote Assignment on Grid Quorum with Association (BVAGQ-AR) dicadangkan bagi mengendalikan replikasi untuk data terbahagi dalam persekitaran pangkalan data teragih dengan menggunakan kos komunikasi dan masa memproses yang rendah untuk satu transaksi. Ciri-ciri utama BVAGQ-AR ialah teknik ini menggabungkan teknik replikasi dan teknik perlombongan data yang membolehkan pengekstrakan pengetahuan yang bermakna dari set data yang besar. Teknik BVAGQ-AR terdiri daripada langkah-langkah berikut. Langkah pertama ialah menggunakan teknik perlombongan data dengan melaksanakan algoritma Apriori dari Association Rules. Ia digunakan untuk mencari perkaitan antara data. Untuk langkah kedua, pangkalan data dipecahkan berdasarkan keputusan analisis perlombongan data. Teknik ini di kaji untuk memastikan replikasi data dapat dilakukan dengan berkesan dalam masa yang sama dapat menjimatkan kos. Kemudian, pangkalan data yang dihasilkan selepas proses pembahagian data diletakkan di pelayan yang berkenaan. Akhir sekali, selepas proses perletakan data, setiap pelayan mempunyai fail pangkalan data tersendiri dan bersedia untuk menjalankan sebarang transaksi dan proses replikasi. Akhirnya, keputusan eksperimen menunjukkan bahawa BVAGQ-AR dapat memastikan konsistensi data dengan kos komunikasi dan masa pemprosesan untuk transaksi terendah berbanding dengan teknik BCSA, PRA, ROWA, HRS and BRS.

ABSTRACT

One of the biggest challenges that data grids users have to face today relates to the improvement of the data management. Organizations need to provide current data to users who may be geographically remote and to handle a volume of requests of data distributed around multiple sites in distributed environment. Therefore, the storage, availability, and consistency are important issues to be addressed to allow efficient and safe data access from many different sites. One way to effectively cope with these challenges is to rely on the replication technique. Replication is a useful technique for distributed database systems. Through this technique, a data can be accessed from multiple locations. Thus, replication increases data availability and accessibility to users. When one site fails, user still can access the same data at another site. Techniques such as Read-One-Write-All (ROWA), Hierarchical Replication Scheme (HRS) and Branch Replication Scheme (BRS) are the popular techniques being used for replication and data management. However, these techniques have its weaknesses in terms of communication costs that is the total replication servers needed to replicate the data. Furthermore, these techniques also do not consider the correlation between data during the fragmentation process. The knowledge about data correlation can be extracted from historical data using techniques of the data mining field. Without proper strategies, replication increases job execution time. In this research, the some-data-to-some-sites scheme called Binary Vote Assignment on Grid Quorum with Association (BVAGQ-AR) is proposed to manage replication for meaningful fragmented data in distributed database environment with low communication cost and processing time for a transaction. The main feature of BVAGQ-AR is that the technique integrates replication and data mining technique allowing meaningful extraction of knowledge from large data sets. Performance of the BVAGQ-AR technique comprised the following steps. First step is mining the data by using Apriori algorithm from Association Rules. It is used to discover the correlation between data. For the second step, the database is fragmented based on the data mining analysis results. This technique is executed to make sure data replication can be effectively done while saving cost. Then, the databases that are resulted after the fragmentation process are allocated at their assigned sites. Finally, after allocation process, each site has a database file and ready for any transaction and replication process. Finally, the result of the experiments shows that BVAGQ-AR can preserve the data consistency with the lowest communication cost and processing time for a transaction as compared to BCSA, PRA, ROWA, HRS and BRS.

DECLARATION

TITLE PAGE

ACKNOWLEDGEMENTS iv

ABSTRAK v

ABSTRACT vi

TABLE OF CONTENT vii

LIST OF TABLES xi

LIST OF FIGURES xii

LIST OF ABBREVIATIONS xiii

CHAPTER 1 INTRODUCTION 1

1.1 Introduction 1

1.2 Problem Statement 3

1.3 Objectives of Research 5

1.4 Scopes of Research 5

1.5 Organization of Thesis 6

CHAPTER 2 LITERATURE REVIEW 7

2.1 Introduction 7

2.2 Theoretical Background 7

2.2.1 Data Grid 7

2.2.2 Data Mining in Grid 9

2.2.3 Data Mining Types 11

2.2.4 Mining Frequent Patterns in Association Rules 14

2.2.5 Database Fragmentation 15

2.3 Literature Review 18

2.3.1 General Data Replication in Grid 19

2.3.2 Data Replication with Data Mining Techniques 20

2.3.3 Data Replication without Data Mining Techniques 26

2.4 Discussion 32

2.5 Summary 38

CHAPTER 3 METHODOLOGY 39

3.1 Introduction 39

3.2 BVAGQ-AR Overview 39

3.3 BVAGQ-AR Transaction Manager (BTM) 43

3.3.1 Data Mining in BVAGQ-AR 45

3.3.2 Data Replication in BVAGQ-AR 54

3.4 The Coordinating Algorithm for Primary Replica 57

3.4.1 Initial Lock 57

3.4.2 Propagate Lock 57

3.4.3 Primary Replica Processing 62

3.5 The Cooperative Algorithm for Neighbour Replica 65

3.6 Replica Management 66

3.6.1 Case 1 - A Set Of Transaction V_η Request to Update Instant b at One Site $i \in S(B)$ 67

3.6.2 Case 2 - Different Sets of Transactions Where V_η and V_ψ Request Instant b at Different Site $i \in S(B)$ 67

3.6.3 Case 3 - Different Set of Transactions V_η and V_ψ Request to Update Instant b at Same Site $i \in S(B)$ 68

3.6.4 Case 4 - A Set of Transaction V_η Request to Update an Unavailable Instant c, at Site $i \in S(B)$ 69

3.7 Illustration Examples 70

3.7.1 Case 1 - Database Mining Management 71

3.7.2 Case 2 - Database Replication Transaction Management 72

3.8 Correctness 75

3.9 Summary 76

CHAPTER 4 RESULTS AND DISCUSSION 77

4.1 Introduction 77

4.2 Hardware and Software Specifications 77

4.3 BVAGQ-AR Experimental Result 79

4.3.1 Experiment 1: Mining S to Identify J_l and the $S(B)^e$ for S' 81

4.3.2 Experiment 2: A Transaction $\boldsymbol{V_\eta}$ Request to Update Instant e at Site E 84

4.3.3 Experiment 3: Different Sets of Transactions, V_η and V_ψ Request Instant e at Different Site 89

4.3.4 Experiment 4: : Different Sets of Transactions, V_η and V_ψ Request Instant e at the Same Site 94

4.3.5 Experiment 5: A set of Transaction V_η Request to Update an Unavailable Instant, i at a Site 99

4.4 Result and Discussion 103

4.4.1 Validity Threats 103

4.4.2 Communication Cost Comparison 104

4.4.3 Replication Job Execution Time Comparison 105

4.5 Summary 108

CHAPTER 5 CONCLUSION **109**

5.1 Introduction 109

5.2 Conclusion 109

5.3 Contributions to Knowledge 111

5.4 Future Work 111

REFERENCES **113**

LIST OF TABLES

Table 2.1	Comparative table of replication strategies with data mining	33
Table 2.2	Comparative table of replication strategies in real time environment	37
Table 3.1	Transactional data from MyGrants	47
Table 3.2	Database with binary variable	51
Table 3.3	Target set	56
Table 3.4	BVAGQ-AR Primary-Neighbours Grid Coordination	70
Table 3.5	The BVAGQ-AR PNGC for $S(B_x)$ = {E, B, D, F, H}.	71
Table 3.6	An example of how BVAGQ-AR handle database mining	71
Table 3.7	An example of how BVAGQ-AR handle concurrent transactions	74
Table 4.1	Server component specifications	78
Table 4.2	System development tools specifications	78
Table 4.3	BVAGQ-AR Grid Coordination	80
Table 4.4	Mining S to Identify J_1 and the $S(B)^e$ for S'	82
Table 4.5	Experimental result for one transaction at one site	85
Table 4.6	Experimental result for two transactions update same data at two sites	90
Table 4.7	Experimental result for two transactions at the same server	95
Table 4.8	Experimental result for transaction request to update unavailable data	100
Table 4.9	Communication cost comparison	104
Table 4.10	BVAGQ - AR Improvement in terms of communication cost (%)	105
Table 4.11	Comparison of job execution time for the minimum number of replication servers	106
Table 4.12	Comparison of job execution time for the maximum number of replication servers	107
Table 4.13	BVAGQ - AR Improvement in terms of job execution time (%)	107

LIST OF FIGURES

Figure 2.1	Horizontal fragmentation	16
Figure 2.2	Vertical fragmentation	17
Figure 2.3	Hybrid fragmentation	18
Figure 2.4	PDDRA architecture	23
Figure 2.5	The framework of ROWA-MSTS	27
Figure 2.6	All replicas in HRS update data	28
Figure 2.7	All replicas in BRS update data	29
Figure 2.8	Data Replication in BVAG	31
Figure 3.1	Replica selection strategies based on data mining techniques	40
Figure 3.2	BVAGQ-AR methodology illustration	43
Figure 3.3	Set of frequent 1-itemsets	48
Figure 3.4	Set of frequent 2-itemsets	48
Figure 3.5	Set of frequent 1-itemsets	49
Figure 3.6	Generating frequent itemsets using the Apriori algorithm	52
Figure 3.7	Examples of data replication in BVAGQ-AR	59
Figure 3.8	An assignment B for data file k where $S(B_k)$ = { E,B,D,F,H }	61
Figure 3.9	One set of transaction at one site	67
Figure 3.10	Two set of transactions at two sites	67
Figure 3.11	Two set of transactions at one site	68
Figure 3.12	Two set of transactions at two sites	69
Figure 3.13	An example of BVAGQ-AR transaction processing	72
Figure 4.1	Five replication servers connected to each other	80
Figure 4.2	Execution time for Experiment 1	81
Figure 4.3	Time diagram for Mining S to Identify J_1 and the $S(B)^e$ for S'	83
Figure 4.4	Execution time for Experiment 2	84
Figure 4.5	Time diagram when a set of transaction $\boldsymbol{V_\eta}$ request to update instante at site E	88
Figure 4.6	Execution time for Experiment 3	89
Figure 4.7	Time diagram when different sets of transactions, $\boldsymbol{V_\eta}$ and $\boldsymbol{V_\psi}$ request instant e at different site.	93
Figure 4.8	Execution time for Experiment 4	94
Figure 4.9	Time diagram when different sets of transactions, V_η and V_ψ request instant e at the same site	98

LIST OF ABBREVIATIONS

ADW	Administrator of Data Warehouse
AUFM	Attribute Usage Frequency Matrix (AUFM)
BRS	Branch Replication Scheme
BSCA	Based on Support and Confidence Dynamic Replication Algorithm
BVAGQ-AR	Binary Vote Assignment in Grid Quorum with Association Rule
DDBMS	Distributed Database Management System
DDS	Distributed Database System
GIG	Global Information Grid
GUI	Graphical User Interface
HRS	Hierarchical Replication Scheme
IBM	International Business Machines
I/O	Input/Output
J&J	Johnson & Johnson
LAN	Local Area Network
NASA	National Aeronautics and Space Administration
NRG	Neighbour Replication on Grid
NTM	NRG Transaction Manager
PDDRA	Pre-Fetching Based Dynamic Data Replication Algorithm
PRA	Pre-Fetching Based Dynamic Replication Algorithm
ROWA	Read-One-Write-All
ROWA- MSTS	Read-One-Write-All Monitoring Synchronization Transactions Systems
RSCA	Replication Strategy Based on Clustering Analysis
RSCP	Replication Strategy Based on Maximal Frequent Correlated Pattern
WAN	Wide Area Network

REFERENCES

Al-Ekram, R. and Holt, R. 2010. OSSR: Optimal Single Site Replication. *Parallel and Distributed Processing with Applications (ISPA).* 433 – 441.

Al-Mistarihi, H.E., and Yong, C. 2008. Replica Management in Data Grid. *International Journal of Computer Science and Network Security.* 8: 22 – 32.

Al-Mistarihi, H.H.E., and Yong, C.H. 2009. On Fairness, Optimizing Replica Selection in Data Grids. *IEEE Transactions on Parallel and Distributed Systems.* 20 (8): 1102 – 1111.

Amir, L. A., Helder, M. R., Nooruldeen, A. D., Qaderd, N. 2018. A Data Replication Algorithm for Groups of Files in Data Grids. *Journal of Parallel and Distributed Computing.* 113: 115-126.

Amir, S. S. and Rahmaniab, M. 2018. Systematic Survey of Big Data and Data Mining in Internet of Things. *Computer Networks.* 139: 19-47

Amjad, T., Sher, M., Daud, A. 2012. A Survey of Dynamic Replication Strategies for Improving Data Availability in Data Grids. *Future Generation Computer System.* 28 (2): 337 – 349.

Amza, C., Cox, A., and Zwaenepoel, W. 2003. Conflict-Aware Scheduling for Dynamic Content Applications. *Proceedings of the 4th USENIX Symposium on Internet Technologies and Systems.*

Aouiche, K. Darmont, J. Boussaïd, O. Bentayeb, F. 2005. Automatic Selection of Bitmap Join Indexes in Data Warehouses. *Data Warehousing and Knowledge Discovery.* 64 – 73.

Apers P. M. G. 1988. Data Allocation in Distributed Database Systems. *ACM Transaction on Database System,* 13: 263 – 304.

Asghari, S. and Navimipour, N. J. 2016. Review and Comparison of Meta-Heuristic Algorithms for Service Composition in Cloud Computing. *Majlesi Journal of Multimedia Processing.* 4(4): 28 – 34.

Ashouraie, M. and Navimipour, N. J. 2015. Priority-Based Task Scheduling on Heterogeneous Resources in the Expert Cloud. *Kybernetes.* 44(10) 1455–1471.

Baiao, F., Mattoso, M., Zaverucha, G. 2000. Horizontal Fragmentation in Object DBMS: New issues and Performance Evaluation. *Proceeding of IEEE International Conference on Performance, Computing, and Communications.* 108 – 114.

Barr, M. and Bellatreche, L. 2010. A New Approach Based on Ants for Solving the Problem of Horizontal Fragmentation in Relational Data Warehouses. *Proceeding of International Conference on Machine and Web Intelligence.* 411 – 415.

Beigrezaei, M., Haghighat, A. T., Meybodi M. R., Runiassy, M. 2016. PPRA: A New Pre-Fetching and Prediction Based Replication Algorithm in Data Grid. *Proceeding of International Conference on Computer and Knowledge Engineering*. 257 – 262.

Berman, F., Hey, A. J. G. and Fox, G. C. 2003. Grid Computing: Making the Global Infrastructure a Reality. *Wiley Series in Communications Networking & Distributed Systems.*

Bernstein, P., Hadzilacos, V., and Goodman, N. 1987. Concurrency Control and Recovery in Database Systems. *Addison-Wesley.*

Bernstein,P., and Goodman, N. 1983. The Failure and Recovery Problem for Replicated Databases. *Proceedings of the 2nd Annual ACM symposium on Principles of Distributed Computing,* 114 – 122.

Bhar, S. and Barker, K. 1995. Static Allocation in Distributed Object Base Systems: A Graphical Approach. *Proceeding of International Conference on Information Systems Data Management*. 92 – 114.

Birman, K. P. 1985. Replication and Fault-Tolerance in The ISIS System. *Proceedings of the 10th ACM symposium on Operating Systems Principles*. 79 – 86.

Boru, D., Kliazovich, D., Granelli, F., Bouvry, P., Zomaya. A. 2013. Energy-Efficient Data Replication in Cloud Computing Datacenters. *Proceeding of IEEE Globecom Workshops*. 446 – 451.

Bsoul, M., Al-Khasawneh, A., Eddien Abdallah, E., and Kilani, Y. 2011. Enhanced Fast Spread Replication strategy for Data Grid. *Journal of Network Computer Application*. 34 (2): 575 – 580.

Budiarto, S.N., Tsukamoto, M. 2002. Data Management Issue in Mobile and Peer to Peer Environment. *Data Knowledge Engineering*. 41: 183 – 204.

Buyya, R. and Venugopal. 2005. A Gentle Introduction to Grid Computing and Technologies. *Computer Society of India*. 29(1): 9 – 19.

Carballeira, F.G., Carretero, J., Calderon, A., Garcia, J.D., and Sanchez, L.M. 2007. A Global and Parallel File System for Grids. *Future Generation Computer Systems*. 116 – 122.

Ceri, S. and Pelagatti, G. 1984. Distributed Databases: Principles and Systems. *New York: McGraw-Hill.*

Charrada, B. F., Ounelli, H. and Chettaoui, H. 2010. An Efficient Replication Strategy for Dynamic Data Grids. *Proceedings of IEEE International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*. 50 – 54.

Charrada, B. F., Ounelli, H. and Chettaoui, H. 2010. Dynamic Period vs Static Period in Data Grid Replication. *Proceeding of IEEE International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*. 565 – 568.

Chiregi, M. and Navimipour, N. J. 2016. A New Method for Trust and Reputation Evaluation in the Cloud Environments Using the Recommendations of Opinion Leaders' Entities and Removing The Effect Of Troll Entities. *Computers in Human Behavior*. 60: 280–292.

Connolly, T. and Begg, C. 2015. Database System a Practical Approach to Design, Implementation and Management, *International Edition, Fifth Edition, Pearson Education.*

Cui, Z., Zuo, D., Zhang. Z. 2013. Based on Support and Confidence Dynamic Replication Algorithm in Multi-Tier Data Grid. *Proceeding of International Journal of Computational Information Systems.* 10: 3909 – 3918.

Daudjee, K., and Salem, K. 2004. Lazy Database Replication with Ordering Guarantees. *Proceedings of Data Engineering.* 424 – 435.

Davidson, B., Molina, H. G. and Skeen, D. 1985. Consistency in a Partitioned Network: A Survey. *ACM Computing Surveys.* 17 (3): 341 – 370.

Deris, M.M., Abawajy, J.H., Taniar, D., Mamat, A. 2009. Managing Data using Neighbour Replication on a Triangular-Grid Structure. *International Journal of High Performance Computing and Networking.* 6 (1): 56 – 65.

Deris, M. M., Evans, D. J., Saman, M. Y., Noraziah A. 2003. Binary Vote Assignment on Grid for Efficient Access of Replicated Data. *International Journal of Computer Mathematics,* 80 (12): 1489 – 1498.

Deris, M.M., Mamat, R., Noraziah, A., Suzuri, H.M. 2003. High Service Reliability for Cluster Server Systems. *Proceeding of IEEE International Conference on International Conference on Cluster Computing.* 280–287.

Doraimani, S. 2007. Filecules: A New Granularity for Resource Management in Grids. *Master thesis. University of South Florida, USA.*

Dullmann, D., Hoschek, W., Jaen-Martinez, J. and Segal, B. 2001. Models for Replica Synchronization and Consistency in a Data Grid. *Proceeding of IEEE Symposium on High Performance on Distributed Computing.* 67 – 75.

Foster, C. K. and Tuecke. S. 2001. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International Journal of High Performance Computing Appllications.* 15 (3): 200 – 222.

Foster, I. and Kesselman, C. 1999. The Grid: Blueprint for a New Computing Infrastructure. *Morgan Kaufmann.*

Gao, L., Dahlin, M., Zheng, J., Alvisi, L. and Iyengar, A. 2010. Dual-Quorum: A Highly Available and Consistent Replication System for Edge Services Dependable and Secure Computing. *Proceeding of IEEE Transactions on Dependable and Secure Computing.* 7 (2): 159 – 174.

Gao, M. D., Nayate, A., Zheng, J. and Iyengar, A. 2005. Improving Availability and Performance with Application-Specific Data Replication. *Proceeding of IEEE Transaction on Knowledge and Data Engineering*. 17 (1):106 – 200.

Golding, R. A. 1992. Weak-consistency group communication and membership. *Ph.D. thesis, Computer and Information Sciences Board, University of California, Santa Cruz.*

Grace, R.K. and Manimegalai, R. 2014a. Data Access Prediction and Optimization in Data Grid Using SVM and AHL Classifications. *International Review Computers and Software*. 9(7):1188 – 1194.

Grace, R.K. and Manimegalai, R. 2014b. Dynamic Replica Placement and Selection Strategies in Data Grids A Comprehensive Survey. *Journal of Parallel and Distributed Computing*. 74(2): 2099 – 2108.

Gray, J., Helland, P., O'Neil, P. E., and Shasha, D. 1996. The Dangers of Replication and A Solution. *Proceedings of ACM SIGMOD International Conference on Management of Data*. 173 – 182.

Gullo, F. 2015. From Patterns in Data to Knowledge Discovery: What Data Mining Can Do. *Physics Procedia*. 62: 18 – 22 .

Hamrouni, T., Slimani, S., Charrada, F. B. 2015a. A Critical Survey of Data Grid Replication Strategies Based on Data Mining Techniques. *Proceeding of Procedia Computer Science*. 51: 2779 – 2788.

Hamrouni, T., Slimani, S., Charrada, F. B. 2015b. A Survey Of Dynamic Replication and Replica Selection Strategies Based on Data Mining Techniques in Data Grids. *Proceeding of Engineering Applications of Artificial Intelligence*. 48: 140 – 158.

Han, J., Cheng, H., Xin, D., Yan, X. 2007. Frequent Pattern Mining: Current Status and Future Directions. *Data Mining and Knowledge Discovery*. 15: 55 – 86.

Han, J., Kamber, M., Pei, J. 2011. Data Mining: Concepts and Techniques. *Morgan Kaufmann Publishers*.

Han, J., Kamber, M., Pei, J. 2012. Data Mining: Concepts and Techniques. Reprint. *Morgan Kaufmann Publishers*

Harandi, M. T., Hou, J., Gupta, I. 2011. Transaction with Replication. *Nikita Borisov, University of Illinois*.

Hoschek, W., Jaen-martinez, J., Samar, Stockinger, A., Stockinger, H. 2000. Data Management in An International Data Grid Project. *Proceeding of ACM International Workshop on Grid Computing*. 77 – 90.

Hossein, R., Parand, F., Riahi, D. 2016. Hierarchical simultaneous vertical fragmentation and allocation using modified Bond Energy Algorithm in distributed databases. *Applied Computing and Informatics*.

IBM. 1999. DB2: Replication Guide and Reference. *New Orchard Road, Armonk, NY 10504 (USA). Number SC26-9642-00.*

Jemal, H. A. and Deris, M. M. 2014. Data Replication Approach with Consistency Guarantee for Data Grid. *IEEE Transactions on Computers.* 63: 2975 – 2987.

Jin, H., Cortes, T., and Buyya, R. 2002. High Performance Mass Storage and Parallel I/O: Technologies and Applications. *IEEE Press/Wiley.*

Joseph, J. and Fellenstein, C. 2004. Grid Computing. *Pearson Education.*

Kemme, B. and Alonso, G. 1998. A Suite of Database Replication Protocols Based on Group Communication Primitives. *Proceeding of International Conference on Distributed Computing Systems.* 156 – 163.

Ko, S.Y., Morales, R., Gupta, I. 2007. New Worker-Centric Scheduling Strategies for Data-Intensive Grid Applications. *Proceedings of the International Conference on Middleware.* 121 – 142.

Krishnamurthy, S., Sanders, W. H. and Cukier, M. 2003. An Adaptive Quality of Service Aware Middleware for Replicated Services. *Proceedings of IEEE Transactions on Parallel and Distributed Systems.* 14 (11): 1112 – 1125.

Ladin, R., Liskov, B., Shrira, L. and Ghemawat, S. 1992. Providing High Availability Using Lazy Replication. *Proceedings of ACM Transactions on Computer Systems.* 10 (4): 360 – 391.

Laura C. V., Schuldt, H., Breitbart, Y. and Schek, H. J. 2009. Replicated Data Management in The Grid: The Re:GRIDiT Approach. *Proceedings of ACM Workshop on Data grids for eScience.* 7 – 16.

Linesch, M. 2007. HP.: Grid – Distributed Computing at Scale an Overview of Grid and the Open Grid Forum. *Copyright © Open Grid Forum 2006, 2007. All Right Reserved. GFD-112.*

Ma, H., Schewe, K.D and Kirchberg, M. 2006. A Heuristic Approach to Vertical Fragmentation Incorporating Query Information Databases and Information Systems. *Proceedings of International Baltic Conference on Databases and Information Systems.* 69 – 76.

Ma, J., Liu, W. and Glatard, T. 2013. A Classification of File Placement and Replication Methods On Grids. *Future Generation Computer System.* 29(6): 1395 – 1406.

Milani, B. A. and Navimipour N. J. 2016. A Comprehensive Review of the Data Replication Techniques in the Cloud Environments: Major Trends And Future Directions. *Journal of Network and Computer Applications.* 64: 229 – 238.

Milani, A. S. and Navimipour, N. J. 2016. Load Balancing Mechanisms and Techniques in The Cloud Environments: Systematic Literature Review and Future Trends. *Journal of Network and Computer Applications.* 71: 86 – 98

Milani, B. A. and Navimipour, N. J. 2017. A Systematic Literature Review of the Data Replication Techniques in the Cloud Environments. *Big Data Research.* 10: 1-7

Mistarihi, H.E. and Yong, C. 2008 Replica Management in Data Grid. *International Journal Computer Science Network Security.* 8(6): 22 – 32.

Mokadem, R. and Hameurlain, A. 2015. Data Replication Strategies with Performance Objective in Data Grid Systems: A Survey. *International Journal of Grid and Utility Computing.* 6(1): 30 – 46.

Moore, R., Baru, C., Marciano, R., Rajasekar, A., Wan, M. 1999. Data-Intensive Computing in the Grid: Blueprint for a New Computing Infrastructure. 105 – 29.

Mulk., V. N. and Mohamed, A. M. 2017. A Prediction Based Dynamic Replication Strategy for Data Intensive Applications. *Computers & Electrical Engineering.* 57: 281-293

Naseera, S., and Murthy, K.V.M. 2009. Agent Based Replica Placement in a Data Grid Environment. *Proceedings of International Conference on Computational Intelligence, Communication Systems and Network.* 426 – 430.

Navathe, S. B. and Ra, M. 1989. Vertical Partitioning for Database Design: A Graphical Algorithm. *SIGMOD Record.* 14: 440 – 450.

Navathe, S., Ceri, S., Wiederhold, G. and Dou,J. 1984. Vertical Partitioning Algorithms for Database Design. *Proceedings of ACM Transaction Database System.* 9: 680 – 710.

Navimipour, N. J. and Milani, F. S. 2014. A Comprehensive Study of the Resource Discovery Techniques in Peer-To-Peer Networks. *Peer-To-Peer Networking and Applications.* 8(3): 474–492

Navimipour, N. J., Navin, A. H., Rahmani, A. M., Hosseinzadeh M. 2015. Behavioral Modeling and Automated Verification of a Cloud-Based Framework to Share The Knowledge and Skills Of Human Resources. *Computers in Industry.* 68: 65-77

Nicola, M. and Jarke, M. 2000. Performance Modeling Of Distributed and Replicated Databases. *Proceedings of IEEE Transactions on Knowledge Data Engineering.* 12(4): 645 – 672.

Noraziah, A., Deris, M.M., Ahmed, N.A., Norhayati, R., Saman, M.Y., Zeyed, M.Y. 2007. Preserving Data Consistency through Neighbour Replication on Grid Daemon. *American Journal of Applied Science.* 4(10): 748 – 755.

Noraziah, A., Abdalla, A. N. and Roslina M.S. 2010. Data Replication Using Read-One-Write-All Monitoring Synchronization Transaction Systems in Distributed Environment. *Journal of Computer Science.* 6 (10): 1033 – 1036.

Noraziah, A., Ainul Azila, C.F, Roslina, M.S., Noriyani, M.Z. and Beg, A.H. 2010. Lowest Data Replication Storage of Binary Vote Assignment Data Grid. *Proceedings of International Conference on Networked Digital Technologies.* 466 – 473.

Noraziah, A., Deris, M. M., Norhayati, R., Rabiei, M. and Shuhadah, W.N.W. 2008. Managing Transaction on Grid-Neighbour Replication in Distributed System. *International Journal of Computer Mathematics*. 86 (9): 1 – 10.

Olston, C. and Widom, J. 2000. Offering a Precision-Performance Tradeoff for Aggregation Queries over Replicated Data. *Proceedings of the International Conference on Very Large Data Bases*. 1 – 12.

Oracle Corporation. 1998. 500, Oracl8i Advanced Replication, *Oracle Technical White Paper*.

Pedone, F., Wiesmann, M., Schiper, A., Kemme, B. and Alonso, G. 2000. Understanding Replication in Databases and Distributed Systems. *Proceedings of the International Conference on Distributed Computing Systems*. 464 – 474.

Pérez, J. M., Carballeira, F. G., Carretero, J., Calderón, A., and Fernández, J. 2010. Branch Replication Scheme: A New Model for Data Replication in Large Scale Data Grids. *Future Generation Computer Systems*. 26: 12 – 20.

Plattner, C. and Alonso, G. 2004. Ganymed: Scalable Replication for Transactional Web Applications. *Proceedings of the International Conference on Middleware*. 155 – 174.

Ram, S. and Narasimhan, S. 1994. Database Allocation in A Distributed Environment: Incorporating A Concurrency Control Mechanism and Queuing Costs. *Management Science*. 40: 969 – 983.

Ram, S. and Narasimhan, S. 1995. Incorporating The Majority Consensus Concurrency Control Mechanism into the Database Allocation Problem. *ORSA Journal of Computing*. 7: 244 – 256.

Ranganathan, K. and Foster, I. 2001. Identifying Dynamic Replication Strategies for A High-Performance Data Grid. *Proceedings of the International Grid Computing Workshop, Springer*. 2242: 75 – 86.

Ranganathan, K. and Foster, I. 2001. Identifying Dynamic Replication Strategies for a High-Performance Data Grid. *Proceedings of Second International Workshop on Grid Computing, Springer-Verlag*. 75 – 86.

Ren, K., Li, Z., and Wang C. 2010. LBDRP: A Low-bandwidth Data Replication Protocol on Journal-based Application. *Computer Engineering and Technology*. 89 – 92.

Ritu Garg, R. and Singh, A. K. 2015. Adaptive Workflow Scheduling in Grid Computing Based on Dynamic Resource Availability. *Engineering Science and Technology*. 18(2): 256 – 269.

Sacca, D. and Wiederhold, G. 1985. Database partitioning in a cluster of processors. *Transaction Database System*. 10: 29 – 56.

Sánchez, A., Montes, J., Dubitzky, W., Valdés, J.J., Pérez, M.S., Miguel, P.D. (2008). Data Mining Meets Grid Computing: Time To Dance. In: Data Mining Techniques in Grid Computing Environments. *John Wiley & Sons*. 1 – 16.

Sashi, K., and Thanamani, A. S. 2010. Dynamic Replica Management for Data Grid. *International Journal of Engineering and Technology*. 2 (4).

Sashi, K., and Thanamani, A.S. 2010. A New Replica Creation and Placement Algorithm for Data Grid Environment. *Data Storage and Data Engineering*. 265 – 269.

Sathya,S. S., Kuppuswami, S. and Ragupathi, R. 2006. Replication Strategies for Data Grids. *Proceedings of International Conference on Advanced Computing and Communications*. 123 – 128.

Satio, Y. and Shapiro, M. 2005. Optimistic Replication. *Proceedings of ACM Computing Survey*. 37 (1): 1 – 44.

Schneider, F. B. 1993. Replication Management Using The State-Machine Approach. *Distributed systems (2nd Edtition), ACM Press/Addison-Wesley Publishing.*

Shorfuzzaman, M., Graham, P., and Eskicioglu, R. 2010. Distributed Popularity Based Replica Placement in Data Grid Environments. *Proceedings of International Conference on Parallel and Distributed Computing, Applications and Technologies*. 66 – 77.

Shorfuzzaman, M., Graham, P., and Eskicioglu, R. 2011. QoS-Aware Distributed Replica Placement in Hierarchical Data Grids. *Proceedings of IEEE International Conference on Advanced Information Networking and Applications*. 291 – 299.

Stockinger, H. 2001. Distributed Database Management Systems and the Data Grid. *Proceeding of IEEE Symposium on Mass Storage Systems and Technologies*. 1 – 1.

Souri A. and Navimipour, N. J. 2014. Behavioral Modeling and Formal Verification of a Resource Discovery Approach in Grid Computing. *Expert Systems with Applications*. 41(8): 3831–3849.

Stockinger, H., Samar, A., Allcock, B., Foster, I., Holtman, K., and Tierney, B. 2001. File and Object Replication in Data Grids High Performance Distributed Computing. *Proceedings of IEEE International Symposium*. 76 – 86.

Tackett, J.J., Gunter, D., and Brown, L. 1995. Special Edition Using Linux. *Que Corporation USA*.

Tamhankar, A.M., Ram, S. 1998. Database Fragmentation and Allocation: An Integrated Methodology and Case Study. *Proceeding of IEEE Transactions Systems, Man and Cybernetics, Part A: Systems and Humans,* 28 (3): 288 – 305.

Tang, M., Lee, B.S., Tang, X. and Yeo, C.K. 2006. The Impact on Data Replication on Job Scheduling Performance in the Data Grid. *International Journal of Future Generation of Computer Systems*. 22: 254 – 268.

Terry, D., Demers, A., Petersen, K., Spreitzer, M., Theimer, M. and B.Welsh. 1994. Session Guarantees for Weakly Consistent Replicated Data. *Proceedings of 3rd International Conference on Parallel and Distributed Information Systems.* 140 – 149.

Thomas, R. H. 1979. A Majority Consensus Approach to Concurrency Control for Multiple Copy Databases. *Proceedings of ACM Transactions on Database Systems.* 4 (2): 180 – 209.

Tian, T., Luo, J., Wu, Z., Song. A. 2008. A Pre-Fetching-Based Replication Algorithm in Data Grid. *Proceedings of the 3rd International Conference on Pervasive Computing and Applications.* 526 – 531.

Tu, M. 2006. A Data Management Framework for Secure and Dependable Data Grid. *Ph.D. dissertation. University of Texas, USA.*

Tu, M., Li, P., I-Ling Yen, Thuraisingham, B., and Khan, L. 2010. Secure Data Objects Replication in Data Grid. *Proceedings of IEEE Transactions on Dependable and Secure Computing.* 7 (1): 50 – 64.

Vadim, K. 2018. Overview of Different Approaches to Solving Problems of Data Mining. *Procedia Computer Science.* 123: 234-239

Vazhkudai, S., Schopf, J. 2003. Using Regression Techniques to Predict Large Data Transfers. *International Journal on High Performance Computer Application.* 17(3): 249 – 268.

Wang, J., Wu, H., Wang, R. 2017. A New Reliability Model in Replication-Based Big Data Storage Systems. *Journal of Parallel and Distributed Computing.* 108: 14 – 27.

Wiesmann, M., Pedone, F., Schiper, A., Kemme, B., and Alonso, G. 2000. Database Replication Techniques: A Three Parameter Classification. *Proceedings of IEEE Symposium on Reliable Distributed* Systems. 206 – 215.

Yin, J., Alvisi, L., Dahlin, M. and Lin, C. 1999. Volume Leases for Consistency in Large-Scale Systems. *Proceedings of IEEE Transaction on Knowledge and Data Engineering.* 11 (4): 536 – 576.

Yu, H. and Vahdat, A. 2002. Design and Evaluation of a Conit-Based Continuous Consistency Model for Replicated Services. *Proceedings of ACM Transactions on Computer Systems.* 20 (3): 239 – 282.

Yuan, Y., Wu,Y., Yang, G. and Yu, F. 2007. Dynamic Data Replication based on Local Optimization Principle in Data Grid and Cooperative Computing. *Proceedings of International Conference on Grid and Cooperative Computing.* 815 – 822.

Zaki, M.J. and Meira, W. Jr. 2014. Data Mining and Analysis: Fundamental Concepts and Algorithms. *Cambridge University Press.*

Zhang, C. and Zhang, Z. 2003. Trading Replication Consistency for Performance and Availability: an Adaptive Approach. *Proceedings of the International Conference on Distributed Computing Systems*. 687 – 695.

Zhao, W., Xu, X., Wang, Z., Zhang, Y. and He, H. 2010. A Dynamic Optimal Replication Strategy in Data Grid Environment. *Proceedings of International Conference on Internet Technology and Applications*. 1 – 4.

Zhao, W., Xu, X., Xiong, N., and Wang, Z. 2008. A Weight-Based Dynamic Replica Replacement Strategy in Data Grids. *Proceeding of IEEE on Asia-Pacific Services Computing Conference*. 1544 – 1549.

Zhou, W., Wang L., and Jia, W. 2004. An Analysis of Update Ordering in Distributed Replication Systems. *Future Generation Computer Systems*. 20 (4): 565 – 590.