INTERNATIONAL DEVELOPMENT PLANNING REVIEW
ISSN:1474-6743 | E-ISSN:1478-3401

PERFORMANCE EVOLUTION OF DYNAMIC DATA MANAGEMENT IN GRID COMPUTING

Ms. R. Prema
New Horizon College,Bangalore,India

Abstract:
Data Management is one of the important features of the data grid, where the large amounts of data are distributed geographically all over the world. In general, data is a specialization of grid that is focused on processing large datasets. Huge amounts of data and the worldwide distribution of data stores contribute to the complexity of the data management challenge. Effective data management is an important issue in today's enterprise environment. Data replication technique are much discussed by the data grid researches in order to create the multiple copies of data and places them in various location to minimize the file access time. This paper describes the Replication strategy called Replication Strategy based on Clustering Analysis (RSCA)is proposed. This work collect files based on the recent access habits of the users.So Data Mining (DM) is introduced to find out the correlations.

Keywords: Data grid, Data Replication, Data Mining, file correlation.

I. INTRODUCTION

In data grid, the individual machines communicate and coordinate in order to process a large amount of data efficiently. The members of the data grid are located in different geographical locations and may be clustered together at one or more sites. A grid connects all these locations and enables them to share the data.

An important technique for data management in grid system is the replication technique. Data replication is an optimization technique in order to improve scalability, fault tolerance, high data availability and low bandwidth consumption.

Replication determines which file to be replicated, when the new replica should be created and where the new replicas to be placed. The main aim of using replication is to reduce access latency and network bandwidth consumption. It also improves the reliability by creating the multiple copies of the same data.

On the other side, there has been recently significant interest on using data mining in grids. Data mining is defined as the process of extracting the hidden data, and information from large amounts of data. The data mining grid concept allows data mining process to be deployed in a grid environment. The grid provides indeed an effective computational support for distributed data mining applications Data replication strategies can then benefit from data mining techniques for solving the task like discovering data file correlations, Storage of file based on past history.

To ensure efficient and fast access to such huge and widely distributed data is hindered by the high latencies of the Internet. To address these problems this research work proposed new algorithm called Transmogrified BHR Algorithm which offers high data availability, low bandwidth

consumption, increased fault tolerance, and improved scalability of the overall system this algorithm shows the better performance in a distributed manner. RSCA and MBHR algorithm were analyzed against TBHR to provide dynamic data management in Grid Computing.

II. BACKGROUND AND RELATED WORK

In general the file correlation that is the group of files is connected together and the group of correlated files may be used as granularity for replication. Even though data mining is applied in numerous areas, the application of data mining to replication in data grids is limited. In this paper we focus on replication strategies based on clustering analysis using data mining techniques. It helps how the data mining techniques improves the performance of data grid.

In [1] Sasi and Thanamani proposed, a Modified BHR algorithm to overcome the limitations of the standard BHR algorithm. The performance of the proposed algorithm is improved by minimizing the data access time and avoiding unnecessary replication.

In [2] J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2011.

In [3] Lakshmi and Thanamani proposed a new dynamic data replication strategy, called DMDR, which consider a set of files as granularity. Their strategy gathers files according to a relationship of simultaneous accesses between files by jobs and stores correlated files at the same site.

In [4] A. K. Kayyoor, A. Deshpande, and S. Khuller. Data placement and replica selection for improving co-location in distributed environments. Computing Research Repository (CoRR), 2013, the combined problem of data placement and replication, given a query workload, to minimize the total resource consumption and by proxy, the total energy consumption, in very large distributed or multi-site read-only data stores.

In [5] S. Y. Ko, R. Morales, and I. Gupta. New worker-centric scheduling strategies for data-intensive grid applications. In Proceedings of the 8th ACM/IFIP/USENIX International Conference on Middleware, Newport Beach, CA, USA, pages 121–142, 2007 proposed various metrics, both deterministic and randomized, that can be used with worker-centric scheduling and found that metrics considering the number of file transfers generally give better performance over metrics considering the overlap between a task and a storage.

In [6] N. Saadat and A. M. Rahmani. PDDRA: A new pre-fetching based dynamic data replication algorithm in data grids. Future Generation Computer Systems, 28(4):666–681, 2012., members in a VO (Virtual Organization) have similar interests in files. Based on this assumption and also file access history, PDDRA predicts future needs of grid sites and pre-fetches a sequence of files to the requester grid site, so the next time that this site needs a file, it will be locally available. This will considerably reduce access latency, response time and bandwidth consumption.

In [7] J. Jiang, H. Ji, G. Xu, and X. Wei. ARRA: an associated replica replacement algorithm based on Apriori approach for data intensive jobs in data grid. Key Engineering Materials, 439-440:1409– 1414, 2010, the data mining techniques can be applied to access historical data of data grids and how do they infer file correlations knowledge and use them to enhance replication strategies performance.

III. REPLICATION STRATEGY BASED ON DATA MINING APPROACH

The association rules, frequent sequence mining are mainly used to identify the file correlation in data mining techniques.

Replication Strategy based on Clustering analysis (RSCA)

Replication Strategy is based on the existence of the correlations among the data files accessed according to the access history of the users. At the first stage, a clustering analysis is conducted on the file access history of all client nodes in the grid over a period of time. The outputs of this operation are correlated file sets related to the access habits of users. At the second stage, replication is done on the basis of those sets, which achieves the aim of pre-fetching and buffering data. The clustering method adopted is used to group into equivalence classes all the files that are similar according to a given equivalence relation. The set of files in the same equivalence class are called correlative file sets.

-Adopted data mining technique:

A given strategy can exploit several kinds of knowledge, extracted through a data mining technique in order to perform a given task, like association between file attributes, clustering grid sites into disjoint groups, etc.

-Data used in the data mining process:

A given data mining process uses as input data extracted from the data grid on which the strategy is applied

-Explicit knowledge:

Explicit Knowledge consists in the extracted patterns from the data mining algorithm, like frequent sequences, association rules, clusters, etc.

-Data mining periodicity:

The data mining algorithm is triggered at each file request or at each period.

-Centralized/Decentralized data mining:

The centralized data mining gathers all data into the central site. Then the data mining algorithm runs on the data. In the decentralized case, it is based on fundamentally distributed algorithms that do not require the centralization of data and other resources.

IV. FORMS OF DATA PRE-PROCESSING

- Data Cleaning:

The first step is Data Cleaning. It removes all the noisy data, incomplete data and inconsistent data.

-Data integration:

All the information is combined to perform analysis. This helps in improving the efficiency and speed of the data mining process.

-Data Reduction:

This technique helps in sorting and obtaining only relevant data. It focuses on reducing the number of attributes and original data volume.

- Data transformation:

The data is aggregated and converted here.

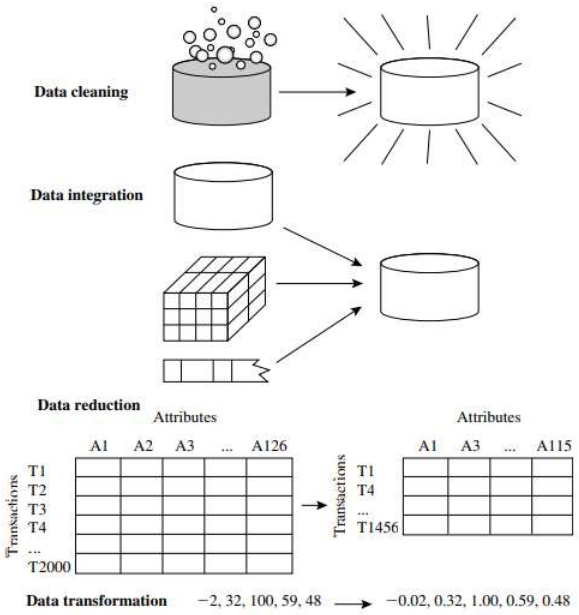


Fig1 Forms of Data Preprocessing

V. BENEFITS OF DATA REPLICATION STRATEGIES

- Availability: The replication strategy is that, it provides maximum availability. So replication is the better way to improve the availability of data in an distributed environments.

-Reliability: Data replication is the process of storing the same data in multiple locations to improve data availability and accessibility, and to improve system resilience and reliability.

-Scalability: Another important metric of replication is scalability. The scalability is more dependent on model than replication algorithm.

-Adaptability: The replication algorithm must be adaptive to provide support to all nodes present in a data grid at any given time.

-Performance: When the availability of data increases then the performance of the data grid environment also increases.

VI. DIFFERENT PARAMETERS AND THEIR IMPORTANCE

All the replications strategies tries to reduce the access latency, thereby reducing the job response time and also increase the performance of the data grids. If the requested data is very near to the site, so that data can be accessed efficiently. It helps in increasing the performance of the system and also provides better response time. If more number of replicas is in a node, the cost of maintaining them becomes an overhead for the system. The storage is utilized in an optimal way and the cost of replica maintenance is minimized.

Some of the parameters are:

- Reduced access latency.
- Reduced bandwidth consumption.
- Less maintenance cost.
- Strategic replica placement.
- Effective network usage

- Job execution time.
- Increased fault tolerance.

VII.DESCRIPTION OF TBHR ALGORITHM

Initially, the user submits a job to the grid. The data are produced in a master site, then master site distributes data to each other region header. The jobs are assigned to computing elements, the places where the jobs are executed. When the job needs the data, and it is not present in the local storage, replication takes place. The replicated files are not stored in all the requested sites. Instead, the file is stored in the site where the file is accessed for the maximum time, with assumption that files recently accessed by a client are likely to be accessed by nearby clients and the files accessed recently are likely to be accessed again.

Algorithm: TBHR Region Based Algorithm

Inputs : Grid Topology, Scheduled job and bandwidth details

Outputs : Mean job Execution Time, Average Storage Used, Network Usage, Number of replications, shortest path, Characteristic Path Length.

Methods:

1. if (Requested File not in Local Site)

   fetch from the nearby site within the region

2. Create Cluster on file accessing history in the grid over a period of time

3. Proceed to replicate among the correlated file sets, which is related to the access habits of users

TBHR Algorithm:

```
if (free space available in SE)
Store it;
Else {
    if (duplication if replica in other sited within region)
    Terminate optimizer;
Else {
     Sort files in SE using LFU
```

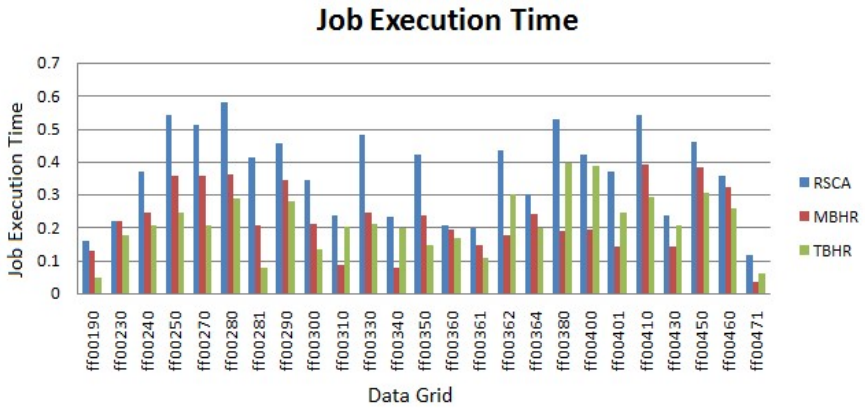
```
For (each file in SE)
{
if file is duplicated in other sites within region
Delete it;
if (of the enough space to store new Replica)
Break;
}}
For (each file in sorted list)
 {
   if (access frequency of new Replica > access frequently of the file
   Delete file
   if (enough free space)
   break
 }}
            if (enough free space)
            Store new replica
           }
```

VIII.EXPERIMENTAL RESULTS AND ANALYSIS

The *Transmogrified Bandwidth Hierarchy Replication Algorithm* was experimented with the IBM Log Dataset.Our proposed algorithm with IBM Log dataset is implemented in MATLAB and proved that our algorithm improves better performance.

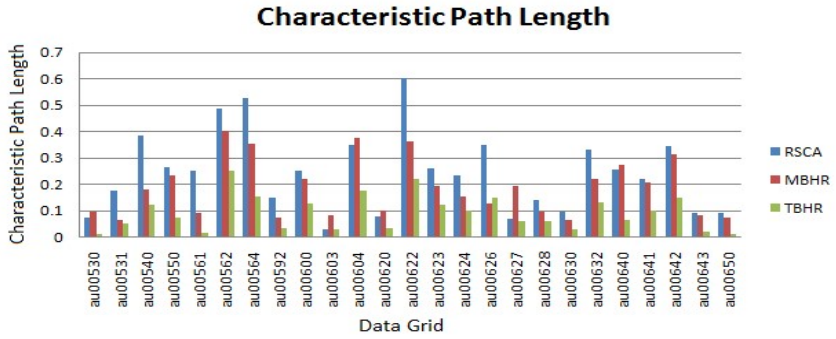
Job Execution Time:

Graph shown below depicts various levels of Job Execution Time on SONAR IBM Log Dataset over RSCA algorithm, MBHR algorithm and TBHR algorithm. The graph clearly shows that TBHR algorithm performs better compared with RSCA algorithm and MBHR algorithm. Also this graph clearly indicates RSCA algorithm has a very low performance compared with both MBHR and TBHR algorithm.



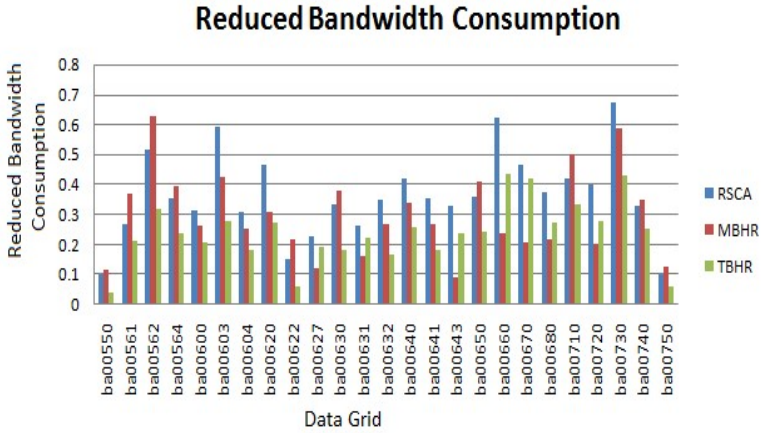
Characteristic Path length:

Graph shown below depicts various levels of Characteristic Path Length on SONAR IBM Log Dataset over RSCA algorithm, MBHR algorithm and TBHR algorithm. The graph clearly shows that TBHR algorithm performs better compared with RSCA algorithm and MBHR algorithm.



Reduced Bandwidth Consumption:

Graph shown below depicts various levels of Reduced Bandwidth Consumption on SONAR IBM Log Dataset over RSCA algorithm, MBHR algorithm and TBHR algorithm. The graph clearly shows that TBHR algorithm performs better compared with RSCA algorithm and MBHR algorithm



IX.CONCLUSION AND FUTURE WORK

Our proposed algorithm, TBHR minimized execution time, saved storage size, reduced bandwidth consumption, improves performance on characteristic path distance and improves efficient network usage.

In future work it can be combined with scheduling to achieve much better performance. The dynamism of the sites can be included as an area of future work in that sites can join and quit the grid at any time. In future this model can be deployed in a real grid environment. The proposed Algorithm increases the data availability by making dynamic replica creation. It also reduces the unnecessary replication. It places the replica in a appropriate location so as to reduce the placement cost.

References:

[1] In Sashi, Thanamani, “Dynamic replication in a data grid using a Modified BHR Region Based Algorithm”, *Future Generation Computer Systems*, Elsevier, 27(2011), pp.202-210

[2]J. H. Jiang et al., "ARRA: An Associated Replica Replacement Algorithm Based on Apriori Approach for Data Intensive Jobs in Data Grid", Key Engineering Materials, Vols. 439-440, pp. 1409-1414, 2010.

[3]Lakshmi, Thanamani, “Performance Evolution of Dynamic Replication in a Data Grid using DMDR Algorithm”, International Journal of Engineering Research & Technology, ISSN: 2278-0181, Vol. 5, Issue. 10, 2016, pp. 389-394.

[4]] S. Goel and R. Buyya, "Data replication strategies in wide-area distributed systems," in Enterprise service computing: from concept to deployment, ed: IGI Global, 2007, pp. 211-241. [5] O. Wolfson, S. Jajodia, and Y. Huang, "An adaptive data replication algorithm," ACM Transactions on Database Systems (TODS), vol. 22, pp. 255-314, 1997. International Journal of Computer Sciences and Engineering Vol.6 (5), May 2018, E-ISSN: 2347-2693 © 2018, IJCSE All Rights Reserved 963

[6] S. Ghemawat, H. Gobioff, and S.-T. Leung, The Google file system vol. 37: ACM, 2003.

[7] R. M. Rahman, K. Barker, and R. Alhajj, "Replica placement design with static optimality and dynamic maintainability," in Cluster Computing and the Grid, 2006. CCGRID 06. Sixth IEEE International Symposium on, 2006, pp. 4 pp.-437.

[8] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on, 2010, pp. 1-10.

[9] U. Čibej, B. Slivnik, and B. Robič, "The complexity of static data replication in data grids," Parallel Computing, vol. 31, pp. 900- 912, 2005.

[10] T. Loukopoulos and I. Ahmad, "Static and adaptive distributed data replication using genetic algorithms," Journal of Parallel and Distributed Computing, vol. 64, pp. 1270-1285, 2004.

[11] H. Lamehamedi, Z. Shunt, B. Szymanski, and E. Deelman, "Simulation of dynamic data replication strategies in data grids," in Parallel and Distributed Processing Symposium, 2003. Proceedings. International, 2003, p. 10 pp.

[12] R.-S. Chang and H.-P. Chang, "A dynamic data replication strategy using access-weights in data grids," The Journal of Supercomputing, vol. 45, pp. 277-295, 2008.

[13] S. Acharya and S. B. Zdonik, "An efficient scheme for dynamic data replication," 1993.
[14] H. Huang, W. Hung, and K. G. Shin, "FS2: dynamic data replication in free disk space for improving disk performance and energy consumption," in ACM SIGOPS Operating Systems Review, 2005, pp. 263-276.
[15] S.-M. Park, J.-H. Kim, Y.-B. Ko and W.-S. Yoon, "Dynamic data grid replication strategy based on Internet hierarchy," in International Conference on Grid and Cooperative Computing, 2003, pp. 838-846.
[16] W. Li, Y. Yang, and D. Yuan, "A novel cost-effective dynamic data replication strategy for reliability in cloud data centers," in IEEE ninth international conference on Dependable, autonomic and secure computing (DASC), 2011, pp. 496-502.
[17] N. Saadat and A. M. Rahmani, "PDDRA: A new pre-fetching based dynamic data replication algorithm in data grids," Future Generation Computer Systems, vol. 28, pp. 666-681, 2012.
[18] X. Sun, J. Zheng, Q. Liu, and Y. Liu, "Dynamic data replication based on access cost in distributed systems," in Fourth International Conference on Computer Sciences and Convergence Information Technology, 2009. ICCIT'09. , 2009, pp. 829-834.
[19] M. Tang, B.-S. Lee, X. Tang, and C.-K. Yeo, "The impact of data replication on job scheduling performance in the Data Grid," Future Generation Computer Systems, vol. 22, pp. 254-268, 2006.
[20] S.-Q. Long, Y.-L. Zhao, and W. Chen, "MORM: A Multi-objective Optimized Replication Management strategy for cloud storage cluster," Journal of Systems Architecture, vol. 60, pp. 234-244, 2014.
[21] A. Doğan, "A study on performance of dynamic file replication algorithms for real-time file access in data grids," Future Generation Computer Systems, vol. 25, pp. 829-839, 2009.
[22] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in IEEE 26th symposium on Mass storage systems and technologies (MSST), 2010, pp. 1-10.
[23] Q. Wei, B. Veeravalli, B. Gong, L. Zeng, and D. Feng, "CDRM: A cost-effective dynamic replication management scheme for cloud storage cluster," in IEEE International Conference on Cluster Computing (CLUSTER), 2010, pp. 188-196.
[24] M. Lei, S. V. Vrbsky, and X. Hong, "An on-line replication strategy to increase availability in data grids," Future Generation Computer Systems, vol. 24, pp. 85-98, 2008.
[25] R. M. Rahman, K. Barker, and R. Alhajj, "Replica placement design with static optimality and dynamic maintainability," in
[26]Leyli Mohammad Khanli, AyazIsazadeh et al., “PHFS: A dynamic replication method, to decrease access latency in the multi-tier data grid”, Future Generation Computer Systems, Elsevier, 27(2011), pp.233-244.
[27]NaimeMansouri, “A Predication-Based Replication Algorithm for Improving Data Availability in Grid Environment”, Journal of Telecommunication, Electronic and Computer Engineering, ISSN: 2180-1843, Vol. 6, No. 1,Jan-June 2014.

[28]NazaninSaadat, Amir MasoudRahmani, "PDDRA: A New pre-fetching based dynamic data replication algorithm in data grids", ELSEVIER, Future Generation Computer Systems, 28(2012), pp.666-681.
[29]Sang-Min Park, Jai-Hoon Kim, Young-BaeKo, Won-Sik Yoon, "Dynamic Data Replication Strategy Based on Internet Hierarchy BHR", in: Lecture notes in Computer Science Publisher, vol. 3033, Springer-Verlag, Heidelberg, 2004, pp. 838-846.
[30]SarraSlimani, TarekHamrouni et al., "New Replication strategy based on maximal frequent correlated pattern mining for data grids", IEEE, International Conference on Parallel and Distributed Computing, Applications and Technologies, pp.144-151.
[31]TianTian, JunzhouLuo et al., "A Prefetching-based Replication Algorithm in Data Grid", Third International Conference on Pervasive Computing and Applications, IEEE.
[32]Zhongqiang, Decheng et al., "Based on Support and Confidence Dynamic Replication Algorithm in Multi-tie Data Grid", Journal of Computational Information Systems, 10(2013), pp. 3909-3918.
[33] Gui Liu, HaiLiang Wei et al., "Research on Data Interoperability Based on Clustering Analysis in Data Grid", International Conference on Interoperability for Enterprise Software and Applications China, IEEE, ISBN: 978-0-7695-3652-1.