ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

Data Replication System in Cloud based on Data Mining Techniques

Vijay. G. R[1], Dr. A. Rama Mohan Reddy[2]

PhD Scholar, Department of Computer Science & Engg, JNTUA, Anantapur, AndraPradesh, India[1]

Professor, Dept. of CSE, SVU College of Engineering, Tirupati, AndraPradesh, India[2]

Abstract*:* Replication is a widely used method for achieving high availability in database systems. The data replication approach can be used for solutions such as Sharing data among remote offices, Sharing data among dispersed users, Make server data more accessible, Distribute solution updates, Back up data, Provide Internet or Intranet Replication. The main aim of this research is to provide a better technique for solving the drawbacks that currently exist in the literary works of data replica method in cloud environment. Here, we intend to propose Data Replication System based on data mining techniques. In proposed method, we will use a cloud computing with Combination of replication algorithm and job scheduling policy for data replication in the data cloud environment through monitoring all job process. The data replication will be done by identifying the frequently used data patterns in the large database of a node. This will be done by frequent pattern mining algorithm. The system availability and replica will be measured and identifying the location in which the replicated data will be stored.

Keywords: Data Replication, Pattern Mining, Association Rules

I. INTRODUCTION

Database Replication is the frequent electronic copying of data from a database in one computer or server to a database in another so that all users share the same level of information. The result is a distributed database in which users can access data relevant to their tasks without interfering with the work of others. However data replication is a fascinating topic for both theory and practice. On the theoretical side, many strong results constraint what can be done in terms of consistency: e.g., the impossibility of reaching consensus in asynchronous systems the blocking nature of CAP theorem, and the need for choosing a suitable correctness criterion among the many possible .On the practical side, data replication plays a key role in a wide range of contexts like caching, back-up, high availability, wide area content distribution, increasing scalability, parallel processing, etc. Finding a replication solution that is suitable in as many such contexts as possible remains an open challenge [1], [3].

Data replication has been widely used to improve the performance of data access in traditional wired/wireless networks. With data replication, users can access the data without the support of network infrastructure, and can reduce the traffic load [13].

Having multiple replicas of a database improves availability since transactions can continue to be executed by other replicas should one go down. Furthermore, bringing a crashed server backup following a failure can be simplified by copying a replica's state instead of rebuilding the crashed server's state from logs [14]. The replication mechanism determines which file should be replicated, when to create new replicas and where the new replicas should be placed. Replication methods can be classified as static and dynamic [19]. Replication components can therefore be subjected to realistic large scale loads in a variety of scenarios, including fault-injection, while at the same time providing global observation and control [18].Replication is a cost effective way to increase availability and used for both performance and fault tolerant purposes thereby introducing a constant trade-off between consistency and efficiency.

Replication is the most providing way for travelling salespeople and roaming disconnected users and enables mobile users with laptops to be updated with current database information when they connect and to upload data to a server. Data is generated and then replicated [2]. Active and passive Replication are two types of replication techniques. In active replication all replicas receive and process the same sequence of client requests. In Passive replication the clients send their requests to a primary, which executes the requests and sends update messages to the backups [5] .The goal of replication is to shorten the data access not only for user accesses but enhancing the job execution performance [15]. Replication provides both improved performance and reliability for mobile computers by creating multiple replicas of important data [16].

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

Three main requirements of database replication are the performance, the availability and the consistency of data. These requirements are in conflict with each other because a change for the benefit of one of the criterion implies a change (minimization) at the expense of the other criteria. Performance and Availability, Network Load Reduction are the most common reasons for using replication [4]. Database replication can be performed in at least three different ways such as Snapshot replication, that functions by periodically sending data in bulk format, Merging replication which allows various sites to work autonomously and at a later time merge updates into a single, uniform result, Transactional replication where the replication agent monitors the server for changes to the database and transmits those changes to the other backup servers, Statement based replication that intercepts every SQL query and sends it to different replicas and each replica (server) operates independently [6].

Data mining, a relatively young and interdisciplinary field of computer science, is the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The goal of data mining is to extract knowledge from a data set in a human understandable structure and involves database and data management, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of found structure, visualization and online updating. The large set of evolving and distributed data can be handled efficiently by Incremental Data mining, Parallel Data mining and Distributed Data Mining [7].

The Advantages of the data replication are, Clients at the site to which the data is replicated experience improved performance because those clients can access data locally rather than connecting to a remote database server over a network. Another advantage is Clients at all sites experience improved availability of replicated data. If the local copy of the replicated data is unavailable, clients can still access the remote copy of the data. These advantages do not come without a cost. Data replication is also used to improve performance [17].The Disadvantage of data replication is that an update to any given logical object must be propagated to all stored copies of that object. A difficulty that arises immediately is that some sites holding a copy of the object might be unavailable (because of a site or network failure) at the time of the update [2].

Further the data replication can also be used in certain other systems like Byzantine Fault Tolerant where the system allows for concurrent transaction execution without relying on a centralized component that is essential for having both performance and robustness [8] ,Job scheduling performance in hierarchical data grid where data replication is used to reduce job execution time by reducing job data access time [20], Three-tier Distributed Systems in which clients (the client-tier) interact with a middle tier (the mid-tier) that forwards client requests to server replicas (the end-tier) [9], Improving the Performance of Cluster Architecture where data replication managed to improve the reliability of the web cluster system.[10], Wireless networks where the process is splitted into two parts. The first one describes the replication with mobile clients, which disconnect after getting the data. Next mobile units permanently stay online [11]. The mobile unit is allowed to locally replicate shared data and to operate on this data while it is disconnected. The local updates can be propagated to the rest of the system on reconnection [12].

The rest of the paper is organized as follows. Section II explains about the design strategy and the proposed method. Section III shows the result and discussion of our proposed method and finally section IV concludes our proposed method for Data Replication System in clouds based on data mining technique.

II.PROPOSED METHODOLOGY FOR DATA REPLICATION SYSTEM IN CLOUD

Data cloud is to share and analyze the data resources, storage resources and others in a wide network which is dynamic, heterogeneous and distributive. The data distributed across a grid must be available and accessible to several applications with a reasonable performance. It mainly focuses on analyzing massive data. In order to analyze the dynamic, real-time and online data, the whole data cloud system must be improved in order to enhance the access speed and reliability and safety and system's load balance. Therefore, how to choose the replication strategy for data cloud is particular important. Creating replica is to reduce access latency and bandwidth consumption, in other words, it is to reduce the average job execution time and improve the usage of cloud resources. These best data replication algorithms are based on some kind's historical data access information and metadata. This information is in static condition. But, cloud is in dynamic environment. In cloud environment data replicate on a particular node. A data replica is best for some node at a certain point of time is not necessary to be best replication for another node at some different time. Because, workload, CPU capacity, changes in networks, etc. So, to select best data replication algorithm is still a problem. The main aim of this research is to provide a better technique for solving the drawbacks that currently exist in the literary works of data replica method in cloud environment. Here, we intend to propose Data Replication system based on data mining techniques. In proposed method, we will use a cloud computing. Combination of replication algorithm and job scheduling policy for data replication in the data cloud environment through monitoring all job process. The data replication will be done by identifying the frequently used data patterns in the large database of a node. This will be done by frequent pattern mining algorithm. The system availability and replica will be measured and identifying the location in which the

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

replicated data will be stored. The replication will be stored based on the data failure probability and system availability. The popularity or frequency of the data will generate an adaptive threshold value for replication. The replicated data will be equally distributed among the nodes for easy access.

A. System Architecture for Data Replication in a Cloud

Three main requirements of database replication are the performance, the availability and the consistency of data. These requirements are in conflict with each other because a change for the benefit of one of the criterion implies a change (minimization) at the expense of the other criteria. The access to a replicated entity is typically uniform with access to a single, non-replicated entity. The replication itself should be transparent to an external user. In addition, in a failure scenario, a failover of replicas is hidden as much as possible. The data replication in the cloud can be explained with the help of the architectural diagram as shown in figure 1. The architecture shows that it contains three major section like User, Scheduling manager and Replica manager. The users are normally the clients those who access the cloud from different locations. Each user can access the cloud independently and will provide different data access which has the property of replication. The particular task of each user is first given to the scheduling manager divides the task to the corresponding data centers through the replica manager based on the number of user using the particular data center to access the file without collusion. The dynamic data replication strategy consists of three different stages.

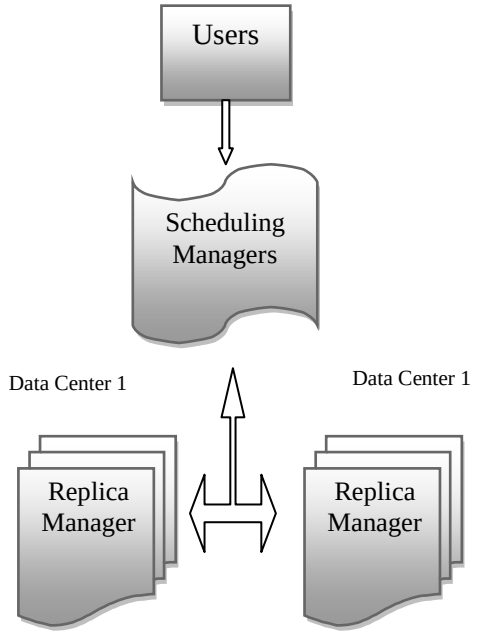

Fig 1: Architectural Diagram of Replication in Cloud Computing.

The first stage is to identify which data file should be replicated and when to replicate in cloud computing to reduce the waiting time. We modify the popularity degree in the first stage of the D2RS algorithm. The modification of the popularity degree is based on double exponential moving average function. The second stage of the D2RS algorithm is to find the required number of replicas and the third stage is to find where to place the new replica data files.

B. Frequent Pattern Mining Algorithm for Data Replication

Mining frequent patterns is generally one of the most important concepts in data mining. A lot of other data mining tasks and theories stem from this concept. The main steps in the frequent pattern mining are to find patterns (item set, sequence, structure, etc.) that occur frequently in a data set. The frequent data mining can support the data replication process to a higher extend.Let D be the set of data items. Let a set $S = \{d_1, d_2 ..... d_n\}$ which belongs to D is called an item set.

The support of an item set S in transaction database T is the number of transaction in the cover of S in T ,

$$\sup(S,T) = |\mathrm{cov}(S,T)|$$

where $\sup()and\ \mathrm{cov}()$ support of item set and cover of item set respectively, The frequency of an item set S in T is the probability of S occurring in a transaction $A \in T$,

$$freq(S,T) = F(S) = \frac{\sup(S,T)}{|T|}$$

An item set is called frequent if its support is no less than a given absolute minimal support threshold σ_{ms} with $0 \le \sigma_{ms} \le |T|$ .The collection of frequent item set in T with respect to σ is denoted by,

$$F_c(T,\sigma) = (S \subseteq D) \mid \sup(S,T) \ge \sigma\}$$

An association rule is an expression of the form $P \Rightarrow Q$, where P and Q are item sets, and $P \cap Q = \{\ \}$. Such a rule expresses the association that if a transaction contains all items in P , then that transaction also contains all items in Q . P Is called the body or antecedent, and Q is called the head or consequent of the rule.

The support of an association rule $P \Rightarrow Q$ in T , is the support of $P \cup Q$ in T , and similarly, the frequency of the rule is the frequency of $P \cup Q$. An, association rule is

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

called frequent if its support (frequency) exceeds a given minimal support (frequency) threshold σ_{ms}. Again, we will only work with the absolute minimal support threshold for association rules.

The confidence or accuracy of an association rule $P \Rightarrow Q$ in T is the conditional probability of having Q contained in a transaction, given that P is contained in that transaction as in the expression given below, $Con(P \Rightarrow Q,T) = \frac{\sup(P \cup Q,T)}{\sup(S,T)}$ The rule is called confident if $Con(P \Rightarrow Q,T)$ exceed given minimal confidence threshold υ with $0 \le \upsilon \le 1$.

C. Association Rule Mining for Pattern Mining

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps. First, minimum support is applied to find all frequent item sets in a database. Second, these frequent item sets and the minimum confidence constraint are used to form rules. While the second step is straightforward, the first step needs more attention. Finding all frequent item sets in a database is difficult since it involves searching all possible item sets (item combinations).

Given a set of items D, a transaction database T over D, and minimal support and confidence thresholds σ_{ms} and υ, the association rule can be found out by $R_a(T,\sigma_{ms},\upsilon)$. Besides the set of all association rules, we are also interested in the support and confidence of each of these rules. Note that the Item set Mining problem is actually a special case of the Association Rule Mining problem. Indeed, if we are given the support and confidence thresholds σ_{ms} and υ, then every frequent item set S also represents the trivial rule $S \Rightarrow \{\ \}$ which holds with 100% confidence. Obviously, the support of the rule equals the support of S. Also note that for every frequent item set D, all rules $P \Rightarrow Q$. Hence, the minimal confidence threshold must be higher than the minimal frequency threshold to be of any effect. Based on the minimum support threshold and confidence threshold values the output differs. The first algorithm proposed to solve the association rule mining problem was divided into two phases. In the first phase, all frequent item sets are generated .The second phase consists of the generation of all frequent and confident association rules. Almost all association rule mining algorithms comply with this two phased strategy. The frequent item set mining of which association rules are a natural extension. Next to the support and confidence measures, a lot of other interestingness measures have been proposed.

III. RESULTS AND DISCUSSION

The proposed method for security model in cloud computing is implemented in the JAVA platform. The cloud data are dived into various datacenters and in each data center the replication data sets are identified and the results we obtain shows that that our method helps in developing an improved data replication system when compare to the existing methods with better efficiency and reduced data loss probability. The frequent pattern mining technique employed in our proposed method provides better replication strategies of data's from the various data centers that are being formed. The data access time in cloud computing plays a major role in its effective functioning and the table 1 given below shows the Data access time taken by our proposed method with different users utilizing the data centers.

Number of Users	Data Access Time (s)		
	Data Center 1	Data Center 2	Data Center 3
5	1.52	1.94	2.03
10	3.15	3.58	3.84
15	5.24	6.87	7.01
20	9.12	9.8	10.5

Table 1: Access time of data by different number of users after performing Data Replication.

As shown in the above table the access time of different set of users using the data from different data centers are being tabulated and for these values the corresponding graphical representation is shown the figure1. The Table 2 shows the similar tabulation value with access time taken by different users when replication strategy is not performed and the values shows that our proposed method of data replication delivers better data access time without much delay.

Table 3 shows the average number of replication files that are created by our proposed method based on the request from the users. These values are compared with that of the existing work and our proposed method proved to be more efficient than the existing one where blind search method is used for replication. The corresponding graphical representation is shown in figure 4.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

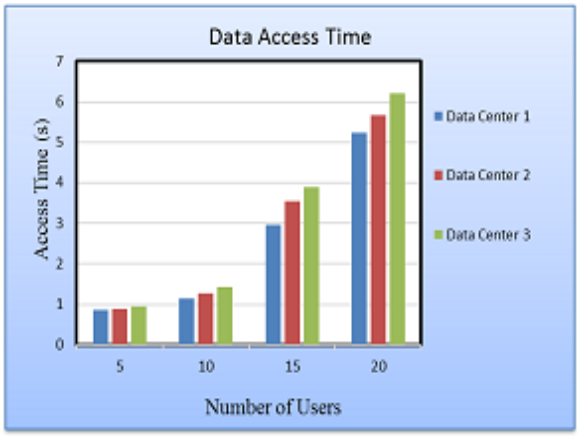

Fig 2: Graphical representation for Data access time in different data centers.

Number of Users	Data Access Time (s)		
	Data center 1	Data center 2	Data center 3
5	0.85	0.89	0.95
10	1.15	1.28	1.43
15	2.95	3.54	3.88
20	5.24	5.67	6.21

Table 2: Access time of data by different number of users without Data replication.

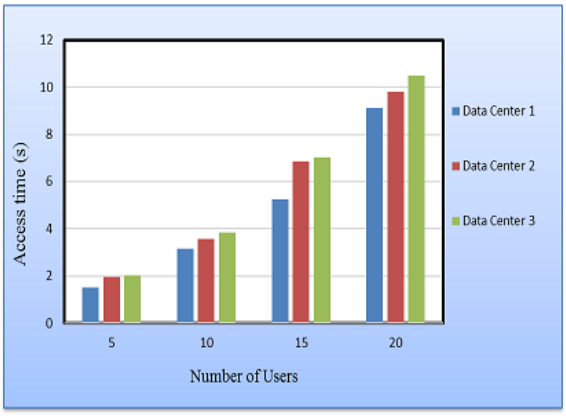

Fig 3: Graphical representation for Data access time in different data centers without replication

Number of Requests	Average Number of Replicated Data Files	
	Proposed Method	Existing Method
100	6.2	5.44
200	5.84	4.95
300	5.2	4.5
400	4.95	4.18
500	4.1	3.7

Table 3: Average Number of Replicated data files corresponding to the user requests.

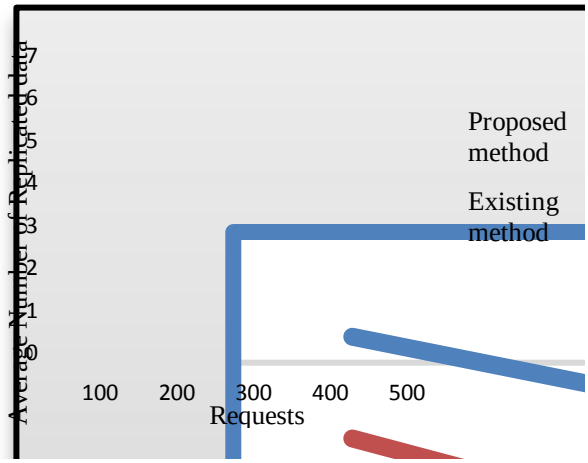

Fig 4: Comparison of Average Number of Replicated data files in proposed and existing methods.

IV. CONCLUSION

In this Paper we have proposed an efficient Data Replication Model for Cloud Computing based on Frequent Pattern Mining. The Data Replication in Cloud Computing requires an in-depth analysis because it is necessary to provide complete access to the users in cloud systems with less time access. The data replication process can ease out the access process of any kind of data without making delay. Our proposed method we utilized the frequent pattern mining algorithm which proved to be a better mechanism to provide efficient replication process to the cloud system. The results we obtain shows that our proposed method has better results when compared with the existing methods of data replication in cloud computing.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 11, November 2013

REFERENCES

[1] Bettina Kemme and Gustavo Alonso , "Database Replication: a Tale of Research across Communities," In.proc.of 36th International Conference on Very Large Data Bases, Vol. 3, No. 1, pp.5-12, Singapore, Sep 2010.

[2] May Mar Oo, Soe and Aye Thida, "Fault Tolerance by Replication of Distributed Database in P2P System using Agent Approach," International Journal of Computers ,Vol .4, No. 1, pp.9-18 , 2010.

[3] Marius Cristian Mazilu, "Database Replication," Database Systems Journal ,Vol. 1, No. 2, pp.33-38, 2010.

[4] Hakik Paci, Elinda Kajo, Igli Tafa and Aleksander Xhuvani, "Adding a new site in an Existing Oracle Multimaster Replication without Quiescing the Replication," International Journal of Database Management Systems (IJDMS), Vol.3, No.3, pp.58-67, Aug 2011.

[5] Sanjay Kumar Tiwari, A. K .Sharma and Vishnu Swaroop , "Issues in Replicated data for Distributed Real-Time Database Systems," (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 ,No.4 ,pp. 1364-1371, 2011.

[6] Salman Abdul Moiz, Sailaja P,Venkataswamy.G and Supriya. N. Pal , "Database Replication: A Survey of Open Source and Commercial Tools," International Journal of Computer Applications ,Vol. 13, No.6,pp.1-8, Jan 2011.

[7] Shashikumar G. Totad, Geeta R. B, Chennupati R Prasanna, N.Krishna Santhosh and PVGD Prasad Reddy , "Scaling Data Mining Algorithms to Large and Distributed Datasets," International Journal of Database Management Systems (IJDMS), Vol.2, No.4,pp.26-35,
Nov 2010.

[8] Rui Garcia, Rodrigo Rodrigues and Nuno Preguica, "Efficient Middleware for Byzantine Fault Tolerant Database Replication," In.Proc. of the sixth conference on Computer systems, Vol.111, pp.107-121, New York, NY, USA, 2007.

[9] Roberto Baldoni, Carlo Marchetti and Sara Tucci Piergiovanni, "Asynchronous Active Replication in Three-tier Distributed Systems,"In.Proc. of Pacific Rim International Symposium on Dependable Computing, pp.19-26, Washington, DC, USA,Dec 2002.

[10] Aznida Hayati Zakaria, Wan Suryani Wan Awang,Zarina Mohamad and Ahmad Nazari Mohd Rose,"Improving the Performance of Cluster Architecture Through Asynchronous Replication Technique," International Journal of Database Theory and Application ,Vol. 4, No. 2, pp.77-86,Jun 2011.

[11] Karol Matiasko and Michal Zabovsky, "A Data Replication For Mobile Environment," Journal of Electrical Engineering, Vol. 59, No. 5,pp.277–280, 2008.

[12] Archana Sharma and Dr.Vineet Kansal, "Replication Management and Optimistic Replication Challenges in Mobile Environment," International Journal of Database Management Systems (IJDMS) ,Vol.3, No.4, pp.81-99, Nov 2011.

[13] Xuejun Zhuo, Qinghua Liy, Wei Gaoy, Guohong Caoy and Yiqi Dai, " Contact Duration Aware Data Replication in Delay Tolerant Networks," In.proc.to IEEE international conference on network protocols, pp. 236-245,2011.

[14] Alexander Thomson and Daniel J. Abadi, "The Case for Determinism in Database Systems," In. Proc. of the Very Large Data Bases Endowment ,Vol. 3, No. 1, Sep 2010.

[15] Nhan Nguyen Dang and Sang Boem Lim, "Combination of Replication and Scheduling in Data Grids," IJCSNS International Journal of Computer Science and Network Security, Vol.7, No.3,pp.304-308, Mar 2007.

[16] David Ratner, Peter Reiher and Gerald J. Popek, "ROAM: A Scalable Replication System for Mobility," Journal of Mobile Networks and Applications archive ,Vol. 9, No 5, pp.537-544, Oct 2004.

[17] R.Sepahvand, A. Horri and Gh. Dastghaibyfard, "Replication and scheduling Methods Based on Prediction in Data Grid," Australian Journal of Basic and Applied Sciences, Vol.5, No.11, pp.1485-1496, 2011.

[18] A. Sousa ,J. Pereira , L. Soares, A. Correia , L. Rocha ,R. Oliveira and F. Moura, "Testing the Dependability and Performance of Group Communication Based Database Replication Protocols,"In .proc. to International Conference on Dependable Systems and Networks, Washington, DC, USA, pp.792 - 801 2005.

[19] Somayeh Abdi and Somayeh Mohamadi , "Two Level Job Scheduling and Data Replication in Data Grid," International Journal of Grid Computing & Applications (IJGCA), Vol.1, No.1, pp.23-37, Sep 2010.

[20] Somayeh Abdi and Somayeh Mohamadi, "The Impact of Data Replication on Job Scheduling Performance in Hierarchical Data Grid," International journal on applications of graph theory in wireless ad hoc networks and sensor networks (GRAPH-HOC), Vol.2, No.3, pp.15-25, Sep 2010.