An Improved Technique for Data Retrieval in Distributed Systems

Ahmed Mateen[1] (*Member, IAENG*), Qingsheng Zhu[1], Salman Afsar[2], Javaria Maqsood[2]

***Abstract* —In current times the world is moving towards the distributed systems, which work with the concept of reliability, availability and performance, as data is stored on multiple sites in distributed manner. If system on one location fails then the data can be accessed from some other storage location, so, the data is easily available in distributed system. In such scenario information retrieval is a big issue, Data might be retrieved by misappropriate user, which might spillover the performance of system as an increase in data retrieval time. In this research work, an improved technique will be designed, by using data replication which is mostly used to manage big volumes of data in a distributed manner. It speeds up data retrieval, reduces data retrieval time and increases data availability, to cope with the issue of data retrieval time. The performance of developed system will be analyzed by giving multiple queries at a time from different systems, which together makes a distributed environment.**

***Index Terms*—Replication, Retrieval Time, Data Availability, Distributed System**

I. INTRODUCTION

In the present condition of appropriated database, the issue is how to deal with an immense measure of information or we can state that how to get the information while reducing retrieval time. Overseeing such an enormous measure of information in unified way is relatively inconceivable because of broadly expanded information. Consequently, quick and viable access to information is exceptionally important. So, information replication is a key element with the goal that we can oversee such a colossal measure of information. An imitated database condition is made number out of database disseminated at various locale. Detailed schema of dataset is used for storage facilitator for the application which require run time data processing [1].

Database system plays key role in almost all the areas. In real world application database system provides numerous services like abstraction and data consistency.

A DBMS provides an interface to users for storage and retrieval. The management system takes care of data accessing concurrently maintaining the integrity of the data. For that distributed data servers are used for integrity and data concurrency. The main focusing point to develop a database is its performance and scalability because when the data size increases its performance response poorly and sometimes system halts or crashes. Performance is an integral and main aspect throughout the DB scenarios [2]. In distributed situations, information replication technique is the methodology of making different information duplicates and putting away the duplicates in various locales. Information replication can enable clients to spare reaction time when errands are being handled in the cloud, and enhance the information accessibility, and lessen the information exchange sums, time and expenses [3].

A. Replication:

Replication is the way toward creating and imitating various duplicates of information at least one locale, and keep the duplicates synchronized so they act as indistinguishable as could be expected under the circumstances. Information implies document, record framework, database et cetera. Replication is a characteristic method for managing disappointments, on the off chance that one replica falls flat, another assumes control. The expression " replica " refers to a site running the database framework programming and storing duplicate of the whole database [4]. The reasons for replication are numerous

B. Server availability

By utilizing replication, information is accessible on numerous destinations and is dodges single purpose of disappointment. When a few destinations are down, the information can be gotten from different sites.

Manuscript received December 15, 2019; revised January 21, 2019. An Improved Technique for Data Retrieval in Distributed Systems.
Ahmed Mateen is with Computer Science Department, Chongqing University China.e-mail: ahmedmatin@hotmail.com
Qingsheng Zhu is with Computer Science Department, Chongqing University China.
Salman Afsar is with Computer Science Department, University of Agriculture Faisalabad, Pakistan.
Javaria Maqsood is with Computer Science Department, University of Agriculture Faisalabad, Pakistan.

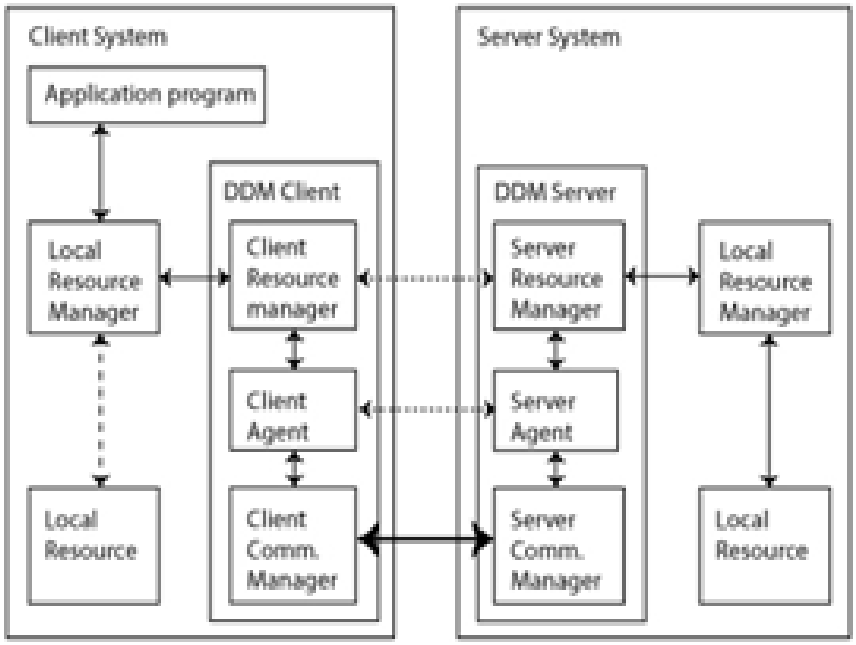


Figure 1: Architectural Model of Distributed Data Base

C. Performance

Replication enables us to find information closer to their passages, subsequently the reaction time will be decreased since the information is nearer.

D. Scalability

As frameworks develop geologically, the quantity of destinations and the quantity of access demand will increase. Data replication is recognized in three various ways.

• Snapshot replication. Data in the principal database is invigorated and duplicated to another

• Consolidating replication. There are no less than two databases which are united into one database.

• Transactional replication. Each one of the customers are outfitted with full type of the fundamental copy of a database and can get changes if any revive in the database is finished.

II. PREVIOUS WORK

Distributed storage frameworks for the most part include excess in putting away substance documents with the end goal that K records are recreated, or deletion coded and put away on N > K hubs. Notwithstanding giving dependability against disappointments, the repetitive duplicates can be utilized to serve a bigger volume of substance get to demand. A request for one of the documents can be either be sent to an efficient hub, or one of the repair gatherings. In this paper, we look to expand the administration limit locale, that is, the arrangement of demand landing rates for the K documents that can be upheld by a coded stockpiling framework. We investigate two parts of this issue: 1) for a given deletion code, how to ideally part approaching solicitations between efficient hubs and repair gatherings, and 2) picking a basic eradication code that boosts the achievable administration limit locale. Specifically, we consider MDS and Simplex codes. In this paper the authors investigation shows that deletion coding makes the framework stronger to skews in record ubiquity rather than just imitating a document at different servers, and that coding and replication together can make the limit locale bigger than either alone [5].

Architectural base information retrieval is based on different architectures like Hadoop architecture. Hadoop architecture is powerful architecture that is designed to explore complex type of data, transform big data and for analyzing of data [6]. In the era of big data Hadoop architecture is used for less retrieval time, huge storage capacity and high availability. In [7], authors describe the approach that can extract information from huge amount of data. In this defined approach uses word count method and inverted index with small testing data in a single environment. This architecture is flexible that allows new users to get benefits of this elaborated architecture. As it is described that this approach uses word count method, so a lot of time will be consumed is this approach. This issue can resolve in future research to identify most frequently used words in data set. In [8], a distributed storage and cluster approach is used for better retrieval of information and efficient storage capacity. In this research a framework is proposed by merging Hadoop base cloud computing

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

platforms and the storage features of Hadoop base disturbed database. This framework is implemented on Linux OS. When there is large number of users, then it will be more efficient to use. The explained framework is based on Hadoop architecture and on Linux platform. Thus, expert people are required to understand this architecture. An idea is proposed [9] in which a web server is designed in map phase using jetty web server that can fast and efficient for searching data in map reduce standard. A searchable mechanism is implemented for real time processing by creating multiple index in web server with the help of multiple search keys and index data node. By using clustering technology, we can handle efficiently traffic and distribute the load on different servers. In future it can be more enhanced by using real time applications for large data sets.

Content based information retrieval retrieves the data on some sort of information like meta data database, image. The model has the capability to narrow down the user's preferences and need and in real life implementation this model is highly effective when data is so large. But for this model metadata is only gathered when user's GPS (global position system) is active to determine the user's exact location and place. In [10] authors proposed a fast image retrieval system that is designed for big data. For image retrieval, firstly all features of image are extracted, and it will take time for large image extraction, so it is required to reduce the feature dimensions for optimizing structure of features. Finally, the similarity matching will determine the retrieved results. The proposed technique for image retrieval works with three main contributions that are feature extraction method, the reason able element ranking and efficient distance metrics that can make better the performance of algorithms. Result shows that the proposed technique can make the more effective performance and retrieved better matching results.

In this proposed technique, feature extraction is very important but still there is need to extract more fractures for more accurate results. The proposed method in [11] improves textual by using map reduce technique. In this map reduce mechanism pattern of text is examined from different data files of big data. The text pattern recognition improves the performance by decreasing number of access to data files of big data. In future, this work can be expanding for distributed system for retrieving texts from different data nodes [12]. Machine learning in information retrieval playing a really important role specially in web search engine, online advertising and recommendation systems. The idea presented in [13] is to accelerate the search operation in big data using neural networks. This cross-correlation technique is used between the user's query and the big data. Moreover, neural networks are used for retrieving of big data even the data is noisy and distorted. The process of information retrieval by using neural network are divided in two parts, first neural networks recognize the input pattern and the second to match it with the given big data. The aim of this work is to manipulate huge amount of data with less time. The proposed model [14] that integrates the scattered data and organize them from multiple heterogeneous data source. This system will retrieve and integrate the non-structured or semi structured data. This model will help the IT business for finding solutions and get information from multiple resources and integrate them. Proposed model is built on Hadoop platform and collect big data using J2EE, the collected data is in the form of XML In [15] authors describe a scheme based on fuzzy similarity for color image retrieval from color library. In image retrieval from big data, color feature is the most important feature. For measuring color similarity in images direct membership value, of histogram from gray levels a gamma histogram plays an important role. For finding membership function from gamma distribution has been proposed. There are several models for these Processes established and practice the basis of implementation. For examples the vector space and probabilistic for information retrieval and 2 position model.

III. METHODOLOG

Data replication, which conveys information documents nearer to the registering VMs, is a viable methodology that lessens the information get to latencies and transmission capacity utilization, along these lines sparing vitality in server farms. There have been a couple of researches that utilize information replication procedures to reduce data get to delay in server farms. Nonetheless, every one of them plan heuristic calculations that don't offer any execution ensure. Therefore, it isn't clear that how execution change can be accomplished all the time with those heuristic calculations [16].

Replication is the way toward replicating client information and a few items (like perspectives and put away methodology) starting with one database then onto the next. The database can exist on the same or distinctive servers. Contingent upon the sort of replication that you select, information can be changed on the duplicate of the database, and afterward resynchronized with the source database [17]. Replication bolsters this development with a worthy response time. Replication is the process of sharing information between different users and to maintain

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

consistency at the same time. Redundant resources share hardware component and software component to make communication more reliable, maintaining fault tolerance and increasing accessibility. In the basic replication model, clients, do not know about the no. of replica exist in the whole communication system. Data replication is applied when the set of nodes in the distributed system are free to communicate with each other. Database replication can be termed as creating and maintaining the duplicate copies of data items in distributed database system [11, 17].

TABLE 1: PARAMETERS ADDRESSED BY TYPE

		Learning base				Targeted data				Extra features	
Type		Classification	Clustering	Data analytics	NLP	Texts	Visuals	Audio	Hybrid	Load balancing	Security
Architectural basis	Information retrieval using hadoob	√	√	√	-	√	√	-	-	√	-
	Hadoob architecture for storage	√	√	√	-	-	-	-	-	√	-
Content basis	Meta data database	√	-	-	-	√	√	√	-	-	√
	Image retrieval	√	-	-	-	√	√	√	√	-	-
	Textual retrieval	√	-	-	-	√	√	√	√	-	-
Machine learning	Neural networks	√	√	√	√	√	√	-	-	-	-
	Probabilistic models	√	-	√	√	√	√	√	-	-	-
	Symbolic learning and rule induction	√	√	-	√	√	√	√	-	-	-
	Evolution based models	-	√	√	√	√	√	√	-	-	-
	Analytic learning and frizzy logic	√	√	√	√	√	-	-	-	-	-
Indexed base	Tree pattern framework	√	-	√		√	-	√	-	√	√
	Artificial intelligence approach	√	√	√	√	√	√	√	-	√	√
	Non-artificial intelligence approach	√	-	√	-	√	√	√	-	√	-
Rule based	Ontology based	√	-	√	-	√	√	-	-	√	-

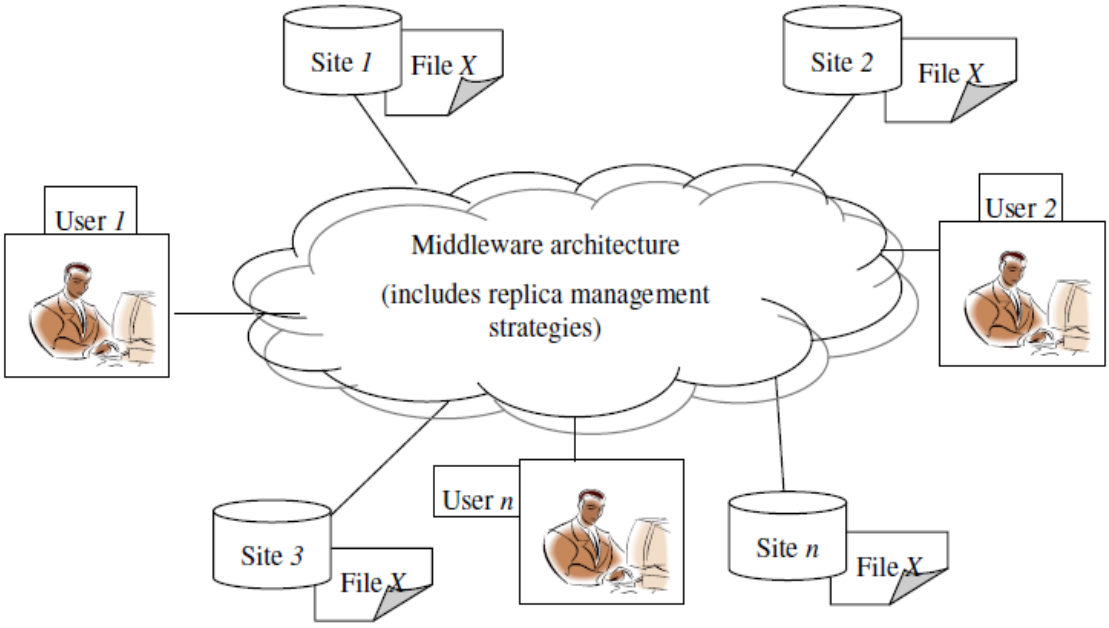


FIGURE 2: OVERVIEW OF REPLICATED FILES AT MULTIPLE SITES

Earlier, the vast majority of the applications were utilizing independent condition where a single centralized server was reacting to numerous clients, working in various areas. To conquer Performance, Availability and Maintenance issues, we can utilize replication arrangement [18]

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

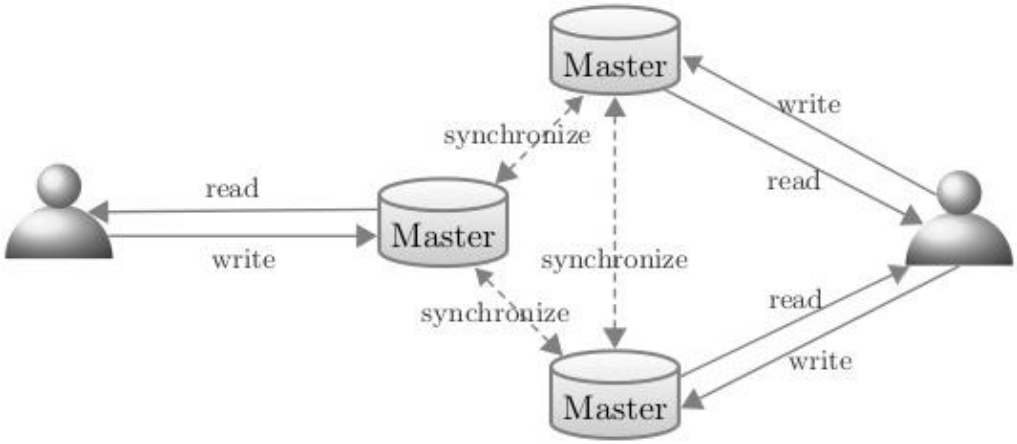


Figure 3: The configuration of replication

```
File  Edit  Format  View  Help
<?xml version="1.0" encoding="utf-8" ?>
<configuration>
  <!--
  Data Source=localhost;Initial Catalog=dengue;Integrated Security=True


  Data Source=192.168.10.4,1433;Network Library=DBMSSOCN;Initial Catalog=checkdata;User ID=smart;Password=smart123;  -->
  <connectionStrings>
    <add name="univote" connectionString="Data Source=192.168.10.5,1433;Network Library=DBMSSOCN;Initial Catalog=dengue;Us
     providerName="System.Data.SqlClient" />
  </connectionStrings>
</configuration>
```

Figure 4: Data Source Configuration.

The chosen configuration for retrieval is simple and efficient being composed from a master and a slave, because it is ideal for distributed systems

A. Configuring the Network

In order to activate a slave system which is not necessarily linked to another system, I connect the system with other through the same IP address. Before I start the data retrieving on slave side, firstly change the connection string in CONFIG File. To create the network, data source must be same both at the master and slave systems.

B. Server configuration

By default, in Express, Developer, and Enterprise Evaluation editions, connection *SQL* by the TCP/IP protocol is disabled. Enable this using SQL Server Configuration Manager. Windows Firewall. While disabling the firewall entirely will work for this component, doing so is not a security best-practice (nor is it required). (Note: in this section, I assume a default configuration. There are many settings that can be changed which affect these steps slightly.) There are two cases depending on the type of SQL Server instance you're connecting to: Default instance (connect by computer name only). Add an allow incoming

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

rule either on TCP port 1433 or the database engine service. Named instance (connect by computer name + instance name). Add an allow incoming rule on UDP port 1434 to access to the SQL Browser service. Add an allow incoming rule on the database engine service.

TABLE 2: ARCHITECTURAL BASE INFORMATION

Type	Category	Subcategory	
Architectural Base	Information retrieval using Hadoop	Word count method	Big data using Hadoop Fault tolerance & availability
	Hadoop architecture for storage	Hadoop cloud computation framework	
	Meta data database	User's location User's current time Common occasion	

C. Implementing the new slave system

To see how fast the data is retrieved at slave system.an application is developed which count the time of data retrieval as we enter the queries length, the data is retrieved.

*Algorithm runs on master database server*Begin
Loop
If transation T arrive from site node Then
Check available server and update avalabile server list
Update salve server satus variable
aa:
If
slaveserverstatus == 0
Then
Check server availability from array list
If server not available Then
Update slaveserver variable status
Goto
aa point
End
if
Retrieve data from master database server
Update slaveserver variable status
Else
if slaveserver status==4
Then
Check server availability from array list
If server not available Then
Update slaveserver variable status
Goto aa point
End
If Retrieve data from slave database server
Update slaveserver variable status=0
Else
Check server availability from array list
If
server not available
Then
Update slaveserver variable status
Goto aa point
End
if
Retrieve data from Slave database server
Update slaveserver variable status (increment)
End if
End loop
End
End if
End loop

IV RESULTS

This experiment uses sets of PC memory distributed databases, the operating system with Windows 2010 database management system with Microsoft SQL Server 2008 R2, and the experimental data of various persons. In the simulation of distributed environment, a comparison between the traditional method and the semi-connection method based on repeat query is carried out, the experimental results as shown

TABLE 3: EXPERIMENTAL RESULTS

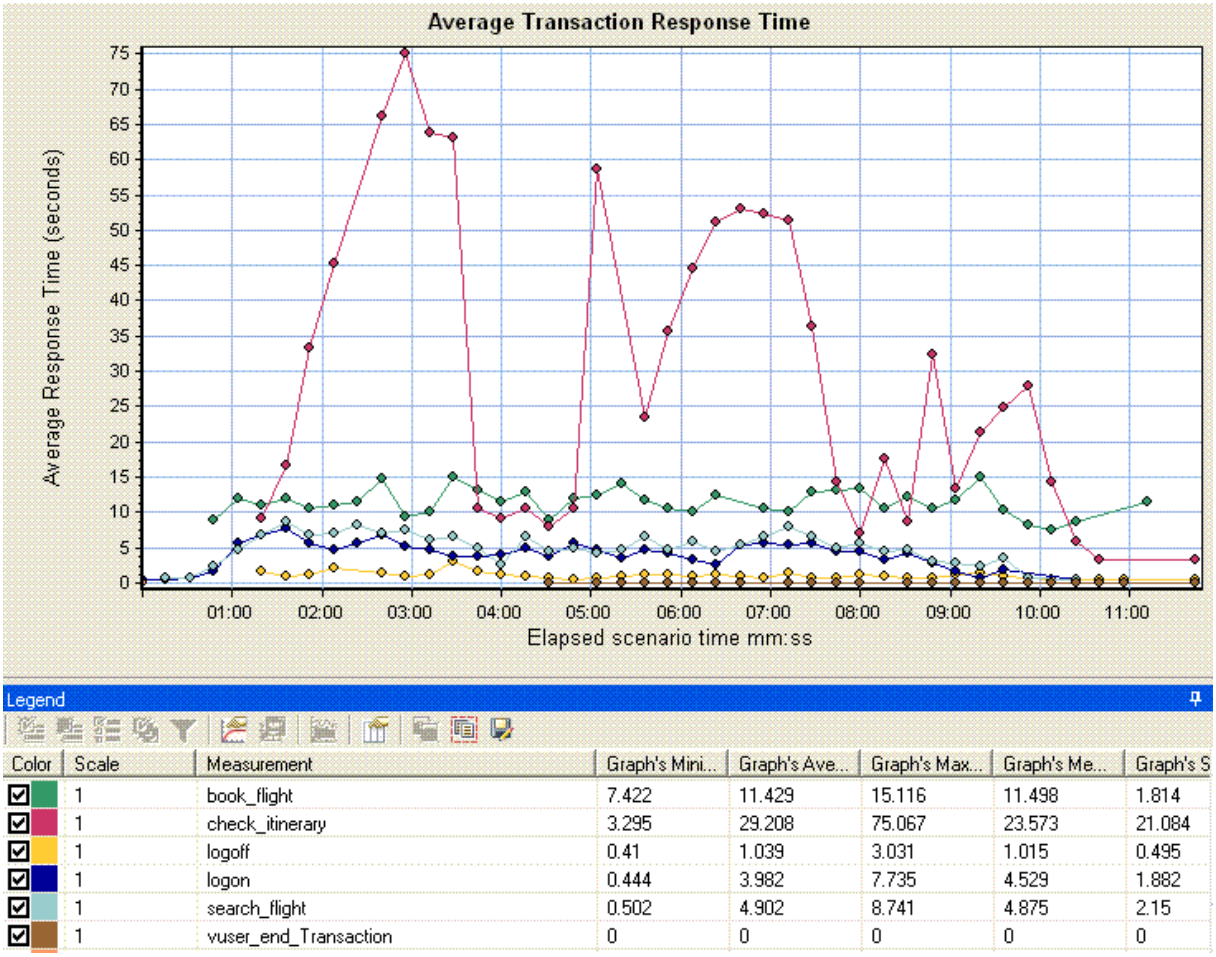
Query scale	Distributed Replication Approach	Without Replication
300	6sec	16sec
1000	19sec	60sec
2000	45sec	175sec
4000	85sec	400sec

Performance Testing and Result Analysis Fig. 1 and Fig. 2 show that at the same time and the same number of concurrent cases, the average response time of transaction before and after optimization of Oracle database.

As you can see, the maximum of the average response time of transaction of Oracle database is 20.355s, the minimum is 0.23s, with an average of 10.185s. The maximum of the Average response time of Transaction of optimized Oracle database is 17.055s, the minimum is 0.23s, with an average of 9.787s.

As you can see, the maximum of number of transactions per second of Oracle database before optimization is 27.125, the minimum is 7.156, with an average of 20.117. The maximum of number of transactions per second of optimized Oracle Database is 24.875, the minimum is 19.094, with an average of 21.234. 1958 Experimental results show that, the change curve of optimized Oracle database is steadier than before optimization. Therefore, Average response time of Transaction of optimized Oracle database is better than before optimization.

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)



Color	Scale	Measurement	Graph's Mini...	Graph's Ave...	Graph's Max...	Graph's Me...	Graph's S
	1	book_flight	7.422	11.429	15.116	11.498	1.814
	1	check_itinerary	3.295	29.208	75.067	23.573	21.084
	1	logoff	0.41	1.039	3.031	1.015	0.495
	1	logon	0.444	3.982	7.735	4.529	1.882
	1	search_flight	0.502	4.902	8.741	4.875	2.15
	1	vuser_end_Transaction	0	0	0	0	0

FIGURE 5: Average Transaction time on database after optimization

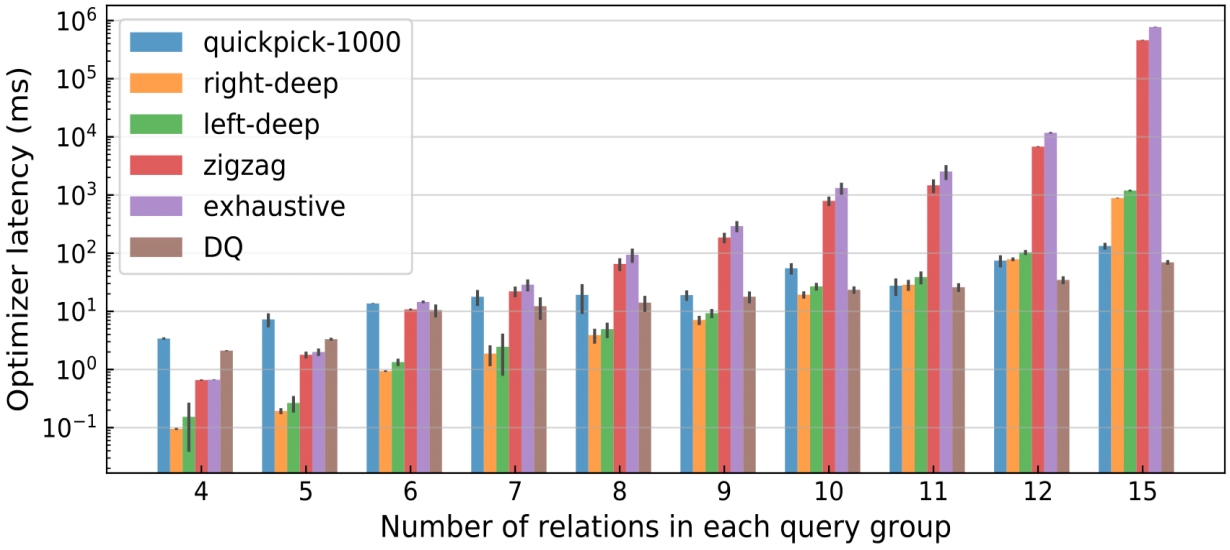


FIGURE 6: Optimized query response time in database

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

IV CONCLUSION

The default B-tree index structure of Oracle database can optimize query performance of massive data. In order to play full role of the index, create the index in the appropriate data table structure, and select a reasonable index column. Full table scan query is inefficient and consumes more system resources, I / O times than the index scan, so indexing is the vital feature during the process of database development. However, unreasonable indexes not only make performance degradation, also lead to the collapse of the database. We can also optimize the query performance of massive data through subregion technology, SQL statements optimization, and operating system.

As the world is moving towards the distributed systems, which work with the concept of reliability, availability and performance, as data is stored on multiple sites in distributed manner, in such situation data retrieval is a big issue. We cope with this scenario and create an application in visual studio which can access multiple queries at a time from master server. The whole retrieval is done by creating multiple copies of data through replication to make the access easy and decrease the retrieval time. The use of SQL server replication technology makes the retrieval process is stable and easy to access data. All the statistical and experimental results are mentioned which can be performed by using the replication technique in SQL Server. Thus, results are also compared with the scenario where retrieval is done without the replication process, and a great difference in retrieval time is observed. In future the same system might be applied by creating a huge network of multiple slave systems, so that it can be used on both small and large network by simply using replication.

REFERENES

[1] Abuqaddom, I. and B. Hammo. 2017. A survey of data replication techniques for MANET databases. Applications of Information Technology in Developing Renewable Energy Processes & Systems (IT-DREPS), 2nd International Conference.

[2] Aktaş, M.,Sarah, E. Anderson, A. Johnston,G. Joshi,S. Kadhe,G. L. Matthews,C. Mayer and E. Soljanin. 2017. On the service capacity region of accessing erasure coded content. Communication, Control, and Computing (Allerton), 55th Annual Allerton Conference.

[3] Bain, T., B. Pavliashvili, J. Sack, M. Benkovich and B. Freeman .2004.Beginning SQL Server 2000,DBA:From Novice to Professional. Apress. 401-468.

[4] Boru, D., D. Kliazovich, F. Granelli, P. Bouvry and Y. A. Zomaya. 2015. Energy-Efficient Data Replication in Cloud Computing Datacenters, Cluster computing. 18(1): 385-402.

[5] Grace, K., H. Maalouf and Z. Mosbeh. 2017. Trust in Real-Time Distributed Database Systems. IEEE. 1:17.

[6] Janpet, J. and Wen. 2013. Reliable and available data replication planning for cloud storage, in Advanced Information Networking and Applications (AINA). IEEE 27th International Conference. 772-779.

[7] Kemme, B., R. Jimènez-Peris, M. Patino-Martinez and G. Alonso. 2010. Database Replication. A Tutorial. Springer. 219-250.

[8] Kumara, A. H. S. and N. R. Sunitha. 2016. A novel archival system with dynamically balanced security, reliability and availability. Computation System and Information Technology for Sustainable Solutions (CSITSS), International Conference. IEEE.

[9] Deshmukh, S., and R. Shah. 2016. Computation Offloading Frameworks in Mobile Cloud Computing: A Survey. In Current Trends in Advanced Computing (ICCTAC), IEEE International Conference, 1-5.

[10] Li, Y., Y. Wu and X. Duan. 2017. Design and implementation of distributed session based on Beansdb. Computational Intelligence and Applications (ICCIA), 2nd IEEE International Conference.

[11] Salman, A. M., P. Sailaja, G. Venkataswamy and S. N. Pal. 2011 "Database Replication: A Survey of Open Source and Commercial Tools". International Journal of Computer Applications 13(6): 0975 – 8887.

[12] Singh, A. K. and U. Shanker. 2016. Database replication techniques: a review.International Journal Of Research Review In Engineering Science & Technology. 5(4).

[13] Tangmankhong, T., P. Siripongwutikorn and T. Achalakul. 2012. Peer-to-peer fault tolerance framework for a grid computing system. Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference.

[14] Vijayakumar, D. , K. G. Srinivasagan and R. S. kumar. 2015. FIR3: A fuzzy inference based reliable replica replacement strategy for cloud Data Centre. Computing and Network Communications (CoCoNet), International Conference. IEEE.

[15]. S. Liu, Q. Qu, L. Chen, L. Ni, SMC: A practical schema for privacy-preserved data sharing over distributed data streams, IEEE Trans. Big Data, 1 (2015) 68-81.[16]. M. Qiu, K. Gai, B. Thuraisingham, L. Tao, H. Zhao, Proactive user-centric secure data scheme using attribute-based semantic access controls for mobile clouds in financial industry, Future Gener. Comput. Syst. (2016) 1

[17]. K. Yu, Y. Gao, P. Zhang, M. Qiu, Design and architecture of dell acceleration appliances for database (DAAD): A practical approach with high availability guaranteed, IEEE, 2015.

[18]. Z. Yan, M. Wang, P. Zhang, A scheme to secure instant community data access based on trust and contexts, IEEE, Xian, China, 2014.

Ahmed Mateen is currently pursuing his Ph.D. Computer Science. From Chongqing university, china. He received the M.Sc. degree from The University of Lahore and M.S. degrees in computer science from the University of Agriculture Faisalabad, Faisalabad, Pakistan. He has an Outstanding Academic Carrier. His research interests are Query Optimization, Modeling and Simulation, Machine Learning, Big Data Analysis and Network Security and Management.

Qingsheng Zhu (M'11) received the B.S., M.S., and Ph.D. degrees in computer science from Chongqing University in 1983, 1986, and 1990, respectively. He is currently a Professor with the College of Computer Science, Chongqing University, and also the Director of the Chongqing Key Laboratory of Software Theory and Technology. His main research interests include Ecommerce, data mining, and service-oriented computing.

ISBN: 978-988-14048-5-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)