Ain Shams University
Faculty of Computer & Information Sciences
Information Systems Department

Dynamic distributed database over cloud environment

A Thesis submitted to Information Systems Department,
Faculty of Computer and Information Sciences, Ain Shams University,
in partial fulfillment of the requirements for
the master degree in Computer and Information Sciences

By

Ahmed Ezzat Abd Alraof Mohamed

Teaching Assistant at Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University.

Under Supervision of

Prof. Dr. Mohammed Fahmy Tolba

Professor, Scientific Computing Department,
Faculty of Computer and Information Sciences,
Ain Shams University.

Prof. Dr. Nagwa Lotfy Badr

Professor, Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University.

Cairo 2016

Acknowledgement

First and foremost, I am grateful to Almighty Allah for His immense blessings and graciously helping me to complete this thesis.

This thesis owes its existence to the help, support, and inspiration of many people. In the first place, I owe my deepest gratitude to my main supervisor Prof. Dr. Mohamed Fahmy Tolba whose sharp sense of research direction have provided invaluable feedback to improve the quality of this thesis. This thesis would not have been possible without his sound advice and encouragement. I would like to express my sincere appreciation and gratitude to my associate supervisor, Prof. Dr. Nagwa Lotfy Badr for her tremendous amount of support, insightful comments, and invaluable assistance.

Last, but definitely not least, I would like to thank all my friends and family members for their endless love and support. My love and heartfelt thank to my parents for their lifelong support in all my endeavors.

Abstract

Distributed Database System (DDS) typically consist of a number of interrelated databases (fragments) located at different geographic sites. These sites can communicate through a network and is managed by a distributed database management system (DDBMS).

The most business organizations need more fixed servers to store their large databases that consist of very large amounts of data, which used by applications at different physical locations. These organizations will charge a lot to design the distributed database infrastructure of their system, especially in the beginning of the work.

Cloud computing allows these organizations to tap into a virtual computing and storage resources over the internet and also benefit from utility like reliability, costs, scalability, as well as pay only for what they use.

However, many emerging applications of distributed database systems generate very dynamic workloads with frequent changes in access patterns from different sites. Consequently, in this realistic dynamic environment, where the access probabilities of nodes to fragments and its replicas change over time, the optimum data re-allocation and replication of those fragments is the only way to increase the performance, efficiency, reliability and availability of the distributed database.

The thesis first address a cluster based distributed and parallel database design over a cloud environment. The proposed architecture and its components are designed for parallel processing the client queries and allow users to access the distributed database from anywhere. It also allows vertical and horizontal fragmentation, allocation and replication decisions to be taken statically at the initial stage of designing the distributed database, without the need of empirical data about query executions. Moreover, it clusters the distributed database sites into disjoints clusters.

Then, a dynamic re-allocation and replication algorithm called optimal fragment reallocation and replication (OFRAR) algorithm was proposed. Which allow migration and/or replication decisions to be taken by each cluster independently of other clusters. This makes it possible to use this algorithm without communication overhead or even using the proposed algorithm on all sites in the system.

Finally,The thesis addresses two types of fragmentation. The first type of the fragmentation is the vertical fragmentation. In this thesis a Full

Vertical Fragmentation, Allocation and Replication (FVFAR) scheme over the cloud environment was presented. The proposed scheme addresses the limitation of the previous vertical fragmentation solutions. It also provides vertical fragmentation, allocation and replication as a service over the cloud.

The second type of fragmentation addressed in this thesis is the horizontal fragmentation. In this thesis an enhanced horizontal fragmentation, allocation and replication algorithm was proposed, which takes the horizontal fragmentation ,allocation and replication decisions at the initial stage of designing the distributed database.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

DDBMS	Distributed Database Management System
DDBS	Distributed Database Systems
OFRAR	Optimal Fragment Reallocation And Replication
MST	Minimum Spanning Tree
VFAR	Vertical Fragmentation, Allocation and Replication
FVFAR	Full Vertical Fragmentation, Allocation and Replication
CRUD	Create, Read, Update and Delete
FAR	Fragmentation, Allocation And Replication
CCR	Communication Cost Range
CDV	Clustering Decision Value
DDBSM	Distributed Database System Manager
ASUM	Attribute Site Usage Matrix
ASM	Attribute Similarity Matrix
AMM	Attribute Manipulate Matrix
ARM	Attribute Read Matrix
FMM	Fragment Manipulate Matrix
FRM	Fragment Read Matrix
FMCM	Fragment Manipulate Cost Matrix
FRCM	Fragment Read Cost Matrix
ALP	Attribute Locality Precedence
MA_F	Number Of Manipulate Accesses For Fragment F
LRA_F	Number Of Local Read Accesses For Fragment F

RRA_F	Number Of Remote Read Accesses For Fragment F
LMA_F	Number Of Local Manipulate Accesses For Fragment F
RMA_F	Number Of Remote Manipulate Accesses For Fragment F
CS_J	Current Size Of Site J In Bytes
RS	Replica Size In Bytes
FS	Fragment Size In Bytes
SC_J	Capacity Of Site J
TA_J	Total Number Of Access To Site J
N	Total Number Of Sites
AC_{FJ}	Access Count For The Fragment F By The Site J
RAC_{FJ}	Read Access Count For The Fragment F By The Site J
$\mathbf{RAC_F}$	Read Access Count For Fragment F
$\mathbf{MAC_F}$	Manipulate Access Count For Fragment F

LIST OF PUBLICATIONS

1. Raouf, A. E. A., Badr, N. L., & Tolba, M. F. (2014, November). Dynamic distributed database over cloud environment. ***Proceedings of International Conference on Advanced Machine Learning Technologies and Applications*** , ***Cairo, Egypt*** (pp. 67-76). **Springer International Publishing**.
2. Raouf, A. E. A., Badr, N. L., & Tolba, M. F. (2015, December). An optimized scheme for vertical fragmentation, allocation and replication of a distributed database. ***Proceedings of 2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)*** , ***Cairo, Egypt*** (pp. 506-513). **IEEE**.
3. Raouf, A. E. A., Badr, N. L., & Tolba, M. F. (2016, October). An Enhanced Distributed Database Design Over the Cloud Environment. ***Proceedings of International Conference on Advanced Intelligent Systems and Informatics***, ***Cairo, Egypt*** (pp. 290-298). **Springer International Publishing**.
4. Raouf, A. E. A., Badr, N. L., & Tolba, M. F. (2016, May). An Enhanced CRUD for Vertical Fragmentation Allocation and Replication Over the Cloud Environment. ***Proceedings of the 10th International Conference on Informatics and Systems***, ***Giza, Egypt*** (pp. 146-152)**. ACM.**
5. Raouf, A. E. A., Badr, N. L., & Tolba, M. F. (2017). Distributed Database System (DSS) Design Over a Cloud Environment. **In *Multimedia Forensics and Security* , Volume 115 of the series Intelligent Systems Reference Library** (pp. 97-116)**. Springer International Publishing.**
6. Raouf, A. E. A., Badr, N. L., & Tolba, M. F. Dynamic Data Re-Allocation and Replication over a Cloud Environment. ***Concurrency and Computation: Practice and Experience***. IF= 0.997, SJR= 0.744. **(submitted).**

Chapter 1

Introduction

1.1 Overview

Cloud computing has become the most essential technology for recent business organizations. It is a distributed environment that uses internet and central remote servers to maintain data and applications. In the case of distributed databases, data storage is located at different dedicated fixed remote servers [2].

These servers can communicate through a network and are managed by a distributed database management system (DDBMS). In addition, each server of the network has autonomous processing capability and also participates in the execution of at least one global database application, which requires accessing data residing at several different sites [3].

Business organizations need more fixed servers to store their large databases that consist of very large amounts of data, which is used by applications at different physical locations.

Cloud computing allows these business organizations to tap into virtual computing and storage resources over the internet and also benefit from utilities like reliability, costs, scalability, as well as pay only for what they use [2].

Fragmentation, replication and allocation are considered the most important design issues that lead to optimal solutions, particularly in a dynamic distributed environment. They also have a great impact on the Distributed Database Systems (DDBS) performance.

Fragmentation is the process of dividing a single database into two or more pieces known as database fragments, such that the combination of the pieces yields the original database without any loss of information [3][4].