



Faculty of Science

A Model for Open Gate Algorithm vs Heterogeneous Distributed Database Issues

A thesis Submitted to

*The Faculty of Science
Ain Shams University*

In partial fulfillment of the requirement for the
degree of doctor of philosophy (PH.D.)
in computer science

By

Abdullah Faisal Atmaz Al-Sebai

Supervised By

Prof. E. Fayed F. M. Ghaleb

*Department of Mathematics
Faculty of Science
Ain Shams University*

Dr. Naglaa M. Reda Taher

*Department of Mathematics
Faculty of Science
Ain Shams University*

Asc. Prof. Hawaf A. H. Telb

*Department of Information Technology
Faculty of Computers and Information
Helwan University*

Cairo Egypt

2010

Acknowledgement

I would like to thank all people, who have helped me to complete my thesis, I am Grateful their assistance.

I am deeply obliged to my supervisors A. Prof. Fayed F. M. Ghaleb and Dr. Naglaa M. Reda for their great help, and their precious time which they have given me.

A great grateful to my family (all of them); they have always encouraged me, I indebted to all; my father, mother, brothers, and sisters.

From my heart big thanks for my friends for their help and encourage.

At the first and the last, a great thanks to my wife who gives me all of her time and support.

Summary

Distributed database systems have become an essential and vital component to trading information and international exchange. Heterogeneous systems spread currently because it allows dealing with various forms of data sources. These systems address the problem of integrating heterogeneous data sources in several ways; the most famous systems use middleware approach. Although many middleware systems have been produced, most of them don't achieve distributed database advantages. Also they face some problems that have bad effects on their performance.

This thesis proposes a new model of a middleware system special for heterogeneous distributed databases. The objective of this system is to achieve some advantages like scalability, autonomy, reliability and high performance.

Also a new wrapper is designed to overcome the main weaknesses in former middleware systems such as bottleneck, and high network communication cost.

The proposed wrapper has been implemented, and tested for many of the various queries for assessing the performance. Also a comparison between the new wrapper's processing time with the time for one of the known systems have been made. Better results and clear improvement in performance have been achieved.

The thesis is divided into five chapters and two appendices:

- Chapter 1: This chapter presents fundamentals of distributed database systems, their basic concepts, advantages and problems, with their different types and architectures. It also gives an overview of the basic technical issues of distributed databases design.
- Chapter 2: This chapter reviews the basic concepts of data integration systems, their definitions, characteristics, classification, and various approaches. It also surveys the most popular systems that rely in its design on middleware systems.
- Chapter 3: In this chapter a detailed study is devoted to a proposed wrapper (IWRAP) for dealing with queries coming from the middleware system. This study includes the proposed wrapper architecture, services and components (*the controller, the schema integrator, and the query translator*). An algorithm with its flow chart for each component is presented. This chapter, also, surveys other wrappers. Finally it presents experimental results of measuring the proposed wrapper's performance compared with another wrapper.

- Chapter 4: This chapter proposes a model for a new middleware system (Open-Gate) which is specially designed for heterogeneous distributed databases using the proposed wrapper. It contains a study of the system's environment, architecture, storages (*the query queue, and the data cache*), and components (*The manager, the global integrator, the query processor, and the data collector*). It introduces algorithms and flow charts for all storages and components. It also focuses on the main characteristics of the proposed system such as autonomy, scalability, reliability and high performance.
- Chapter 5: Conclusion and future work.
- Appendix A: It gives a brief study of some other middleware systems.
- Appendix B: It presents the source code for the proposed wrapper.

Table of Contents

Chapter 1: Distributed Database Systems	1
1.1. Introduction	2
1.1.1. Basic Definitions	2
1.1.2. DDBS Promises	4
1.1.3. DDBS Problems	6
1.2. Types of DDBS	7
1.2.1. Autonomy	7
1.2.2. Homogeneity	8
1.2.3. Distribution	8
1.3. DDBMS Architecture	8
1.3.1. Client-Server	8
1.3.2. Multiple Clients-Multiple Servers	11
1.3.3. Peer-to-Peer	11
1.3.4. Homogeneous and Heterogeneous	11
1.3.5. Parallel vs. Distributed DBMS	13
1.4. Technical Issues of DDB Design	16
1.4.1. Data Distribution	16
1.4.2. Query Processing and Optimization	20
1.4.3. Transaction and Concurrency Control	24
1.4.4. Reliability and Replication Protocols	27
1.4.5. Failures and Recovery	28
Chapter 2: Data Integration Systems	31
2.1. The Concept of Data Integration System	31

2.2. Characteristics of Data Integration	32
2.2.1. Distribution	32
2.2.2. Heterogeneity	35
2.2.3. Autonomy	36
2.3. Approaches of Integration	38
2.3.1. Unstructured Integration	38
2.3.2. Structured Integration	40
2.4. Related Integration Systems	46
2.4.1. MOCHA	46
2.4.2. TSIMMIS	48
2.4.3. Garlic	52
2.4.4. DISCO	54
2.4.5. COIN	55
2.4.6. AMIS	58
Chapter 3: "IWRAP" An Intelligent Wrapper	59
3.1. Related Work	61
3.2. IWRAP Architecture	63
3.3. IWRAP Components	65
3.3.1. The Controller	67
3.3.2. The Schema Integrator	74
3.3.3. The Query Translator	77
3.4. Performance Evaluation	79
Chapter 4: A Model for Open-Gate System	87
4.1. The System Environment	87
4.1.1. Users Interfaces	87
4.1.2. Data Sources	88

4.2. The Proposed Open-Gate Architecture	88
4.3. Open-Gate Storages	89
4.3.1. The Query Queue	91
4.3.2. The Data Cache	92
4.4. Open-Gate Components	97
4.4.1. The Manager	97
4.4.2. The Global Integrator	102
4.4.3. The Query Processor	105
4.4.4. The Data Collector	108
4.5. Open-Gate Characteristics	110
4.5.1. Autonomy	110
4.5.2. Scalability	111
4.5.3. Reliability	112
4.5.4. High Performance	112
Chapter 5: Conclusion and Future Work	115
Appendix A: Samples of Integration Systems	119
Appendix B: IWRAP Source Code	131
List of Publications	145
References	147

List of Figures

1.1	Distributed Database System	4
1.2	Client/Server Architecture	9
1.3	Homogenous DDB	12
1.4	Heterogeneous DDB	13
2.1	Federate Database	41
2.2	Principle Mediator Architecture	44
2.3	Middleware Gateway Database	45
2.4	MOCHA Architecture	47
2.5	TSIMMIS Architecture	50
2.6	Garlic Architecture	53
2.7	DISCO Architecture	54
2.8	COIN Architecture	56
3.1	The Architecture of IWRAP	64
3.2	The Classifier Algorithm	69
3.3	The Classifier Flowchart	70
3.4	The Extractor Algorithm	72
3.5	The Extractor Flowchart	73
3.6	The Schema Integrator Algorithm	75
3.7	The Schema Integrator Flowchart	76
3.8	The Query Translator Algorithm	78
3.9	The Query Translator Flowchart	79
3.10	The Performance of AMIS' Wrapper	82
3.11	The Performance of IWRAP	82

3.12	The Performance Comparison	84
3.13	The Speedup Comparison	84
4.1	Open-Gate Architecture	90
4.2	The Query Queue Algorithm	93
4.3	The Query Queue Flowchart	94
4.4	The Data Cache Algorithm	95
4.5	The Data Cache Flowchart	96
4.6	The Manager Algorithm	99
4.7	The Manager Flowchart	101
4.8	The Global Integrator Algorithm	103
4.9	The Global Integrator Flowchart	104
4.10	The Query Processer Algorithm	106
4.11	The Query Processer Flowchart	107
4.12	The Data Collector Algorithm	108
4.13	The Data Collector Flowchart	109

List of tables

3.1	Abbreviations Summary	66
3.2	Performance Results of AMIS Wrapper	81
3.3	Performance Results of IWRA	81

