

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



Multidatabase Query Optimization

THESIS

Submitted to the Information Systems department, Faculty of
Computer and Information Sciences, Ain Shams University in
partial fulfillment of the requirements for the degree of Master
of Computer and Information sciences

BY

AMIRA REZK ABDOU REZK

B.Sc. Degree in Computer and Information sciences (2000)
Demonstrator, Information System Dept., Faculty of Computer
Science and Information System, Mansoura University.

SUPERVISED BY

Prof. Dr. **MOHAMED ELSHARKAWY**
Faculty of Computer and Information Sciences
Ain Shams University.

Asst. Prof. Dr. **HESHAM ARAFAT**
Faculty of Engineering
Mansoura University.

Cairo – 2006

To my parent ...

Who taught me the value of learning.

Acknowledgment

Thanks are due to ALLAH for getting this work done.

I am very grateful for the encouragement of Professor Esam Khalifa the faculty dean and his support to finish this work.

I wish to express my gratitude to my supervisor Professor Mostafa M. Syaim for his support during the preparation of this work.

I would like to thank my supervisor Professor M. A. Sharkawy for his helpful, guidance and valuable comments on my work.

My sincere thank to my supervisor Assistant Professor H. A. Ali for his cooperative, guidance, and encouragement during the preparation of this work.

Many thanks to my friend Eng. Mervat M. Fahmy for her support and advices.

I am thankful for every one urges me to go forward.

CONTENTES

	ABSTRACT	v
	LIST OF FIGURES	vii
	LIST OF TABLES	ix
1	INTRODUCTION	2
	1.1 Motivation	2
	1.2 Research Goals	3
	1.3 Publications	4
	1.4 Thesis Outline	4
2	MULTIDATABASE SYSTEM ENVIRONMENT.....	7
	2.1 Multidatabase System Why? And What?	8
	2.2 The Multidatabase System Characteristics	10
	2.3 The Multidatabase System Classification	13
	2.4 Multidatabase System vs. Distributed Database System	16
	2.5 Multidatabase systems Architecture	19
	2.6 Multidatabase Systems Problems and Issues	22
	2.6.1 Distribution problems	22
	2.6.2 Heterogeneity problems	22
	2.6.3 Autonomy problems	25
	2.6.4 Database integration Issues	26
	2.6.5 Query processing Issues	26
	2.6.6 Transaction Management Issues	28
	2.7 What is The Next?	29
	2.8 Summary	29

3	QUERY OPTIMIZATION AND RELATED WORK...	31
3.1	Overview Of The Query Optimization	32
3.2	The Rewrite Stage	35
3.2.1	The Representation Step	35
3.2.2	The Transformation Step	35
3.2.3	Applying The Rewrite Stage in The Multidatabase	36
3.3	The Planning Stage	37
3.3.1	Query Evaluation Step	37
3.3.2	Select Access Plan Step	42
3.3.3	Applying The Planning Stage in The Multidatabase	47
3.4	The Enumerate Algorithms	50
3.4.1	Deterministic Algorithms ..	51
3.4.2	Randomized Algorithms	54
3.4.3	Genetic Algorithm	57
3.5	What Is The Next?	59
3.6	Summary	59
4	QUERY OPTIMIZATION FRAMEWORK	61
4.1	The Environment	62
4.1.1	Query Optimization Challenges	62
4.1.2	Problem Definition	63
4.2	The Proposed Framework	64
4.2.1	Schema Integration	64
4.2.2	Database Components Behavior	65
4.2.3	Network challenge	66
4.3	Framework Enhancement	67
4.3.1	Decomposition	67
4.3.2	Optimized Query Plan Generation	68

4.3.3	Execution	69
4.4	The Framework Mechanism	70
4.4.1	The FDBMS Layer Operations	71
4.4.2	The DBCs Layer Operations	73
4.4.3	The Messages	76
4.5	What Is The Next?	78
4.6	Summary	78
5	VALIDATION AND PERFORMANCE EVALUATION ..	80
5.1	The Implementation Consideration	81
5.1.1	Integration	81
5.1.2	The Federated Database Management System ..	82
5.1.3	The Performance Parameters.	83
5.2	The Case Study	84
5.2.1	The Traditional System.	84
5.2.2	The Proposed Framework	85
5.2.3	Illustrated Experiments.	86
5.2.4	Experiments Result	103
5.3	Simulation Results and analysis.	107
5.3.1	Illustrate Experiments.	108
5.3.2	Experiments Result	113
5.4	Evaluation.	121
5.5	Summary	123
6	CONCLUSIONS AND FUTURE WORK	125
	REFERENCES	129
	APPENDIX A	138
	ملخص	١

ABSTRACT

Multidatabase Query Optimization, By: Amira Rezk Abdou
Submitted to the Information Systems department, Faculty of Computer
and Information Sciences, Ain Shams University

The researchers have an interest in the Multidatabase system (MDBS) as a new trend to integrate the pre-existing database systems, as information resources, and provide the global users with a global uniform view. However, the heterogeneity and autonomous of such systems make it difficult to integrate their data.

The most critical issue in Multidatabase system (MDBS) is query optimization, and the most important problem associated to global query optimization in a MDBS is that some required local information about local database components (DBC's) will not be available at the global level due to local autonomy.

The main goal of this work is to solve the problem associated to Query optimization in MDBS. In this thesis an overview of the MDBS characteristics and different issues are introduced. The most previous work that interested in the Multidatabase query optimization are studied carefully. Such study and analysis depict a lot of problems that must be solved to improve the performance of the system such (The lack of information about the local DB in MDBS, Integrate data from heterogeneous sources, Perform large join queries efficiency). The main concern of this work is solving the first one.

The main contribution of this thesis is solving the query optimization problems in the Multidatabase system. To achieve this goal a new framework is proposed to establish a global protocol. This framework can be applied to any autonomy database systems which have the ability to integrate together. It deals with Federated Database System as special case from the Multidatabase system where there is no global catalog.

The proposed framework aims at enhancing the performance of the query optimization process through four designed issues:

- Overcoming the lack of local information.
- Improving the decomposition process.
- Increasing the response of the system.
- Reducing the data transmission in the system.

Validation and evaluation of the efficiency of the proposed framework is done via illustrated experiments of a case study and simulation experimental result. The experiment results indicate that, the proposed framework achieves its goal and objectives to enhance the performance of the query optimization process along the four design issues as follow:

- Collect the local data from its site using the suggested driver overcomes the lack of local information.

- Distributing the decomposition process among the local database component enhances it and avoids wastage of the decomposition effort. The average of eliminating the wastage of the decomposition is around 50% of the submitted Query.

- The suggestion of executing the query partially, i.e. the framework provides the global user with an answer for his/her query even if there are sub-queries that will not be executed, increases the ratio of return result for the user around 45%.

- Applying the proposed routing technique that depends on a set of messages between the two layers reduces the data transmission in the system. The percent of data exchange reducing is increase when the no of join operation which join tables from different DBCs increases.

The simulation profess that the proposed framework achieves its goal to improve the query optimization process and enhance the performance of the Multidatabase system.

Keywords: Multidatabase, Federated database,
Query optimization, Heterogeneity, Autonomy.

LIST OF FIGURES

Figure 2.1	Types of DBS according to distribution, heterogeneity and autonomy.	12
Figure 2.2	Taxonomy of Multidatabase systems	15
Figure 2.3	The ANCI/SPARC architecture of the MDBS	20
Figure 2.4	The ANCI/SPARC architecture of the FDBS	21
Figure 2.5	The most critical problems issues in the Multidatabase system	28
Figure 3.1	The query optimization process	34
Figure 4.1	FDBMS Framework layouts	70
Figure 4.2	FDBMS layer Operations	74
Figure 4.3	DBC's layer Operations	75
Figure 5.1	The integration of the Query2's sub-queries in the traditional system.	88
Figure 5.2	The integration of the Query5's sub-queries in the traditional system.	91
Figure 5.3	The integration of the Query5's sub-queries in the proposed framework	98
Figure 5.4	The integration of the Query5's sub-queries in the second state at the proposed framework.	101
Figure 5.5	The integration of the Query5's sub-queries at the third state in the proposed framework.	103
Figure 5.6	The ratio of executing the Queries in both the proposed framework and the traditional system	103

Figure 5.7	The probability of losing the decomposition effort in the traditional system.	104
Figure 5.8	The size of data exchanged in both the proposed framework and the traditional system	105
Figure 5.9	The Effect of the partial execution in the proposed framework.	106
Figure 5.10	The flow chart of the simulator.	110
Figure 5.11	The relation between the execution of the queries and the no of join belong to the query.	114
Figure 5.12	The probability of losing the decomposition effort in the traditional system.	116
Figure 5.13	The Effect of the partial execution in the proposed framework.	116
Figure 5.14	The relation between the execution of the queries and the no. of join in the query (1 st case)	118
Figure 5.15	The probability of losing the decomposition effort in the traditional system (1 st case)	118
Figure 5.16	The Effect of the partial execution in the proposed framework (1 st case).	119
Figure 5.17	The relation between the execution of the queries and the no. of join in the query (2 nd case)	119
Figure 5.18	The probability of losing the decomposition effort (2 nd case)	120
Figure 5.19	The Effect of the partial execution (2 nd case)	120
Figure A.1	The DBCs integration layout	138

LIST OF TABLES

Table 2.1	Multidatabase system vs. Distributed database system ...8
Table 3.1	Example of selectivity factor 44
Table 3.2	Example of the Cost of join 45
Table 3.3	Example of the Cost of Scans 46
Table 3.4	Summary of search strategies (Enumeration Algorithms) 58
Table 4.1	Exchanged messages descriptions 77
Table 5.1	Comparison between the traditional model and the proposed model 85
Table 5.2	The estimation size of the sub-queries results. 93
Table 5.3	A snapshot of the execution process in example1 ... 111
Table 5.4	The output of the simulator ratio calculation 112
Table A.1	The global Schema of STUDENT 147
Table A.2	The global Schema of COURCE 147
Table A.3	The global Schema of TEACH 147
Table A.4	The global Schema of ENROLL 148
Table A.5	The global Schema of DEPARTMENT 148
Table A.6	The global Schema of MAJOR 148
Table A.7	The global Schema of MINOR 149
Table A.8	The global Schema of FACULTY 149

1

INTRODUCTION

Chapter1

INTRODUCTION

1.1 Motivation

Quick review for the large modern enterprise finds a large number of data sources that contain a huge amount of data. Nowadays, there is a new trend to combine the information from these various systems. Therefore, the enterprise can realize the full value of the data they contain. Nevertheless, these systems are usually not designed to interoperate with each other in a uniform manner [1, 2, 3, 4]. In this way, they are isolated data sources, only used for local purpose. The heterogeneity and autonomous of these systems make it difficult to integrate their data. Throughout the 1980s, the database market matured and companies attempted to standardize on a single database vendor. However, the reality of doing business generally made such a strategy unrealistic [57]. Therefore, there is a growing need for tools to maximize the reusability and interoperability of these arbitrary computing systems. This tool is the Multidatabase management systems (MDBMS). MDBMS is built on the top of the existing database system, to integrate data from them and provide the user with a global uniform view.

This trend debuted commercially in the 1990s under various names. Papers of that time often referred to such products as next-generation gateways, data access middleware and Multidatabase. Then a new terminology Federated database system (FDBS) appears [58].

Although Multidatabase system has the common database systems characteristics, it has its own characteristics that make it more complex than the other types of database system (DBS), and imposes certain constraints that must be taken into consideration. MDBS is an integration of a set of database system components (DBC), each one of these DBCs is located in different location (Distributed), built using various database management system, different hardware platform, different schema, or data model. (Heterogeneous), and work stand-alone, i.e. each DBC is isolated from the other (fully Autonomy). However, when these DBCs decide to integrate together, they give up part of their autonomy. So MDBS can be characterized as Distributed, Heterogeneous and semi-Autonomy [5].

1.2 Research Goals

The most critical issue in Multidatabase system (MDBS) is query optimization. It can be considered the backbone of any successful database system. Therefore, there is a growing need for query optimization algorithm that can effectively deal with Multidatabase system. There are many obstacles and challenges are implied within query optimization in MDBS. These obstacles rise from the nature of query optimization itself, or from the characteristics of the MDBS.

The major challenge of the global query optimization in a Multidatabase system is that some required local information about local database components (DBC) might not be available at the global level due to local autonomy.

In this thesis, the introduced solution is oriented to the federated database system, the special case of the multidatabase system, where there is no global catalog. The goals and research activities are outlined as follows:

- 1- Surveying the Multidatabase system environment, its characteristics, and issues.

- 2- Surveying the query optimization problems in the database system and its challenges in the multidatabase system environment.
- 3- Introducing a framework to manage the relation between the FDBMS and the DBCs, and state a set of rules for any database systems decided to integrate together.
- 4- Providing a solution for the lack of local information based on the introduced framework.

1.3 Publications

- Mostafa Syaim, H. A. Ali, Amira Rezk, "A New Framework For Query Optimization in Multidatabase System Environment". Proceedings of the 14th International Conference on Computer Theory and Applications, Alex, 2004

** Mostafa Syaim, H. A. Ali, Amira Rezk, "A New Framework For Query Optimization in Multidatabase System Environment", MJCSIS Headquarters Journals Staff, Volume 1, Number 0, Jan 2005

- Mostafa Syaim, H. A. Ali, Amira Rezk, Solving Query Optimization Problems in Multidatabase System. Proceedings of 3rd International Conference on Computer Science, Software Engineering, e-Business and Applications, Cairo, 2004

- Mostafa Syaim, H. A. Ali, Amira Rezk, Analysis Of Multidatabase System Framework To Solve Query Optimization Problems. Proceedings of 2nd International Conference on Intelligent Computing & Information Systems, Cairo, 2005.

1.4 Thesis Outline

The work in this thesis is organized as follow:

Chapter 2 introduces some concepts within the Multidatabase system environment contain an answer for the