.

UNIVERSITI PUTRA MALAYSIA

IMPROVING INTEGRITY CONSTRAINTS CHECKING IN DISTRIBUTED DATABASES BY EXPLOITING LOCAL CHECKING

ALI AMER ALWAN

FSKTM 2008 19

IMPROVING INTEGRITY CONSTRAINTS CHECKING IN DISTRIBUTED DATABASES BY EXPLOITING LOCAL CHECKING

By

ALI AMER ALWAN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Science

November 2008

To my Dearest and First Teachers: My Father and Mother
To my lovely brother and sisters
To my lovely wife

Ali

Abstract of thesis presented to Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science

IMPROVING INTEGRITY CONSTRAINTS CHECKING IN DISTRIBUTED DATABASES BY EXPLOITING LOCAL CHECKING

By

ALI AMER ALWAN

November 2008

Chairman : Associate Professor Hamidah Ibrahim, PhD

Faculty : Computer Science and Information Technology

Integrity constraints are important tools and useful for specifying consistent states of a database. Checking integrity constraints has proven to be extremely difficult to implement, particularly in distributed database. The main issue concerning checking the integrity constraints in distributed database system is how to derive a set of integrity tests (simplified forms) that will reduce the amount of data transferred, the amount of data accessed, and the number of sites involved during the constraint checking process. Most of the previous approaches derive integrity tests (simplified forms) from the initial integrity constraints with the sufficiency property, since the sufficient test is known to be cheaper to execute than the complete test as it involved less data to be transferred across the network and always can be evaluated at the target site, i.e. only one site is involved during the checking process thus, achieving local checking. The previous approaches assume that an update operation will be executed at a site where the relation specified in the update operation is located (target site), which is not always true. If the update operation is submitted at a different site, the sufficient test is no longer local as it will definitely access data from the remote sites. Therefore, an approach is needed so that local checking can be performed regardless the location of the submitted update operation.

In this thesis we proposed an approach for checking integrity constraints in a distributed database system by utilizing as much as possible the information stored at the target site.

The proposed constraints simplification approach produces support tests and this is integrated with complete and sufficient tests which are proposed by previous researchers. It uses the initial integrity constraint, the update template, and the other integrity constraints to generate the support tests.

The proposed constraints simplification approach adopted the substitution technique and the absorption rules to derive the tests. Since the constraint simplification approach derives several different types of integrity tests for a given update operation and integrity constraint, therefore a strategy to select the most suitable test is needed. We proposed a model to rank and select the suitable test to be checked based on the properties of the tests, the amount of data transferred across the network, the number of sites participated, and the amount of data accessed.

Three analyses have been performed to evaluate the proposed checking integrity constraints approach. The first analysis shows that applying different types of integrity tests gives different impacts to the performance of the constraint checking, with respect to the amount of data transferred across the network which is considered as the most critical factor that influences the performance of the checking mechanism. Integrating these various types of integrity tests during constraint checking has enhanced the performance of the constraint mechanisms. The second analysis shows that the cost of checking integrity constraints is reduced when various combinations of integrity tests are selected. The third analysis shows that in most cases localizing integrity checking can be achieved regardless of the location where the update operation is executed when various types of integrity tests are considered.

Abstrak tesis yang dikemukakan kepada senat Uinversiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master of Sains

PENAMBAHBAIKAN SEMAKAN KEKANGAN INTEGRITI DALAM PANGKALAN DATA TERAGIH DENGAN MENGEKSPLOITASI SEMAKAN SETEMPAT

Oleh

ALI AMER ALWAN

November 2008

Pengerusi : Profesor Madya Hamidah Ibrahim, PhD

Fakulti : Sains Komputer dan Teknologi Maklumat

kekangan Integriti adalah peralatan yang penting dan berguna untuk menyatakan keadaan konsisten bagi sesuatu pangkalan data. Menyemak kekangan integriti telah terbukti sukar untuk dilaksanakan, khususnya di dalam pangkalan data teragih. Isu utama berkenaan pemeriksanaan kekangan integriti di dalam pangkalan data teragih adalah bagaimana untuk mendapatkan satu set ujian integriti (bentuk teringkas) yang boleh mengurangkan jumlah penghantaran data, jumlah capaian data, dan bilangan lokasi yang terlibat semasa proses semakan kekangan. Kebanyakan pendekatan yang lepas menerbitkan ujian integriti (bentuk teringkas) dari kekangan integriti yang asal dengan ciri yang mencukupi, memandangkan ujian mencukupi berkenaan dikenali lebih murah untuk dilaksanakan berbanding dengan ujian lengkap kerana ia melibatkan kurang data untuk dihantar ke seluruh rangkaian dan seringkali boleh dinilai di lokasi yang dituju, iaitu hanya satu lokasi sahaja yang terlibat semasa proses semakan, oleh itu semakan setempat tercapai. Pendekatan-pendekatan lepas mengandaikan bahawa operasi kemas kini akan dilaksanakan di lokasi di mana hubungan yang dinyatakan dalam

operasi kemas kini ditempatkan (lokasi dituju), yang mana ini tidak selalunya benar. Jika operasi kemas kini tersebut dihantar ke lokasi yang berbeza, ujian mencukupi berkenaan bukan lagi setempat kerana ia pastinya akan mencapai data dari lokasi yang jauh. Oleh itu, satu pendekatan diperlukan agar semakan setempat boleh dilakukan tanpa mengira lokasi operasi kemas kini yang dihantar.

Dalam tesis ini, kami mencadangkan satu pendekatan untuk semakan kekangan integriti di dalam sistem pangkalan data teragih dengan menggunakan sebanyak mungkin maklumat yang disimpan di lokasi yang dituju. Pendekatan mempermudahkan kekangan yang dicadang kan ini menghasilkan ujian sokongan dan ia disatukan dengan ujian lengkap dan ujian mencukupi yang dicadangkan oleh penyelidik-penyelidik terdahulu. Ia menggunakan kekangan integriti yang asal, templat kemas kini, dan kekangan integriti yang lain untuk menjana ujian sokongan.

Pendekatan mempermudahkan kekangan yang dicadangkan mengguna pakai teknik penggantian dan peraturan pengasingan untuk menerbitkan ujian-ujian tersebut. Oleh kerana pendekatan mempermudahkan kekangan menerbitkan beberapa jenis ujian integriti bagi sesuatu operasi kemas kini dan kekangan integriti, maka, satu strategi untuk memilih ujian yang sesuai diperlukan. Kami mencadangkan satu model untuk menyusun dan memilih pengujian yang bersesuaian untuk disemak berdasarkan ciri-ciri ujian, jumlah data yang dihantar ke seluruh rangkaian, bilangan lokasi yang turut serta, dan jumlah data yang dicapai.

Tiga analisis telah dilakukan untuk menilai pendekatan semakan kekangan integriti yang dicadangkan. Analisis pertama menunjukkan penggunaan pelbagai jenis ujian integriti akan memberi kesan yang berlainan kepada prestasi semakan kekangan dari segi jumlah data yang dihantar di dalam rangkaian yang merupakan faktor yang paling kritikal yang mempengaruhi prestasi mekanisma semakan. Menyatukan pelbagai jenis ujian integriti semasa semakan kekangan telah meningkatkan prestasi mekanisma kekangan. Analisis kedua menunjukkan bahawa kos bagi semakan kekangan integriti telah dikurangkan bila pelbagai kombinasi ujian integriti dipilih. Analisis yang ketiga menunjukkan bahawa dalam kebanyakan kes, pemeriksaan integriti secaua setempat boleh dicapai tanpa mengira lokasi di mana operasi kemas kini dilaksanakan apabila pelbagai jenis ujian integriti dipertimbangkan.

ACKNOWLEDGEMENTS

In the name of *ALLAH*, the most merciful and most compassionate. Praise to *ALLAH* S. W. T. who granted me strength, courage, patience and inspirations to complete this work.

I do not think that words can express the gratitude that I have for my supervisor Associate Professor Dr. Hamidah Ibrahim. As a top researcher in database system, her smart intuition on computer science, stimulating suggestions and encouragement helped me in all the aspects of study.

First and foremost, I heartiest would to sincere thanks my supervisor Associate Professor Dr. Hamidah Ibrahim, for her incredible guidance, continuous support, and encouragement. Always having time for me and readily providing her technical expertise throughout the period of my study. I owe more than I can ever repay. Only has the successful completion of this work become possible due to her supervision, she is the first person to thank for making my Master program at the Universiti Putra Malaysia a very enjoyable experience. Her high stand of diplomatic power and professionalism set a great model for me to follow.

To my thesis committee member, Dr. Nor Izura Udzir, I would like to express appreciation for her insightful comments, questions, criticisms, and suggestions on the work. Her critical appraisals of my papers and presentations are extremely valuable for

the improvement in my thinking also her kindness and willingness to help is unforgettable.

I would like to thank many people I have met during my stay in Malaysia for their help, enjoyable discussions and some good times.

This research is partially supported by an IRPA fund number (04-02-04-0797-EA001), which is sponsored by Malaysian Ministry of Science, Technology, and Innovation. Thanks to Universiti Putra Malaysia and Malaysian government for the financial support during my study. Thanks also are due to other members of the academic, and the technical staff in the faculty for their help.

Finally, I would like to express my love and deepest thanks to my noblest father Amer Alwan and my great mother Ebtisam Mohammad were the reason of my success, I'm indebted to them. To my brother, and sisters, who have loved and support me throughout my life, thank you.

Ali Amer Alwan

November 2008

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.

ALI AMER ALWAN

Date:

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AbSIS	Agent-based Semantic Integrity Subsystem
CNF	Conjunctive Normal Form
CPA	Constraint Planning Algorithm
DBMS	Database Management System
DDB	Distributed Database System
DDBMS	Distributed Database Management System
DDL	Data Definition Language
FOL	First Order Logic
RDM	Relation Data Model
SIS	Semantic Integrity Subsystem
SQL	Structure Query Language

CHAPTER 1

INTRODUCTION

1.1 Overview

The validity, accuracy, and semantic of data are significant requirements in modern database applications. Semantic data in database is normally represented under the form of integrity constraints. Integrity constraints are properties, typically depending on the nature of the application domain, which must always be satisfied for the data to be considered *consistent*. Maintaining obedience of data with respect to integrity constraints is an essential requirement, since, if some data lack integrity, then answers to queries cannot be trusted. Databases usually contain massive collections of data that rapidly evolve over time; this makes perfect checking at each update too time consuming a task to be feasible. In this regard, DBMS needs to be extended with the ability to automatically verify that database updates do not introduce any violation of integrity (Christiansen and Martinenghi, 2004; Martinenghi, 2005). The way we pursue here is the so-called *simplification* of integrity constraints. Simplification means to generate a set of integrity tests from the initial constraints whose satisfaction implies the satisfaction of the original constraints in the updated state. The main interest of the simplification process is to obtain a set of integrity tests (simplified forms) that are as easy to evaluate as possible. In this sense, simplification technique is feasible in terms of the cost of evaluating the constraints. *Integrity constraint checking* is the process of ensuring that the integrity constraints are satisfied by the database after it has been

updated. Checking the consistency of a database state will generally involve the execution of integrity tests on the database which verify whether the database is satisfying its constraints or not. The problem of checking integrity constraints in database system has been addressed by many researches, and has been proved to be extremely difficult to implement, particularly in distributed database.

The rest of this chapter is organized as follows. In Section 1.2 the problem statement is explained. Section 1.3 presents the objectives of this research work. The scope of the research is listed in Section 1.4. In Section 1.5 the methodology of this research work is elaborated. The organization of the thesis is outlined in Section 1.6.

1.2 Problem Statement

Maintaining the consistency of database is an important aim in the database system, which means that the data contained in a database is both accurate and valid. Checking the consistency of a database state will generally involve the execution of *integrity tests* on the database which verifies whether the database is satisfying its constraints or not. A naïve approach is to perform the update and then check whether the integrity constraints are satisfied in the new database state. This method, termed *brute force checking*, is very expensive, impractical and can lead to prohibitive processing costs. The brute force strategy of checking constraints is worse in the distributed database context since the checking would typically require data transfer as well as computation leading to complex algorithms to determine the most efficient approach. Allowing an update to execute with the intention of aborting it at commit time in the event of constraint

violation is also inefficient since rollback and recovery must occur at all sites which participated in the update (Feras, 2006; Grefen, 1993; Ibrahim *et al*, 2001; Madiraju and Sunderraman, 2004; Martinenghi and Christiansen, 2005; Madiraju *et al*, 2006).

The main issue concerning checking the integrity constraints in distributed database system is how to derive a set of integrity tests that will reduce the amount of data transferred, the amount of data accessed, and the number of sites participated during the process of constraint checking so that local checking can be achieved. Thus, not surprisingly most of the previous approaches such as Gupta (1994) and Ibrahim (1998) derive simplified form of the initial integrity constraints (integrity tests) with the sufficiency property, since the sufficient test is known to be cheaper to execute than the complete test as it involved less data to be transferred across the network and always can be evaluated at the target site, i.e. only one site will be involved during the checking process. However, the previous approaches assume that an update operation will be executed at a site where the relation specified in the update operation is located, which is not always true. For example, consider a relation *R* which is located at site 1. An insert operation into *R* is assumed to be submitted by a user at site 1 and the sufficient test generated is used to validate the consistency of the database with respect to this update operation, which can be performed locally at site 1. But if the same update operation is submitted at different site, say 2, the sufficient test is no longer appropriate as it will definitely access data from site 1 which is remote to site 2. Thus, the sufficient test which is proven to be local test by previous study is no longer appropriate. Hence, we can conclude that most of the previous works are limited due to this strict assumption. Many of the previous approaches failed to exploit the available information at the target

site and explore the various types of integrity tests to ensure local checking can always be achieved. With respect to this, an approach is needed so that local checking can be performed regardless the location of the submitted update operation.

1.3 Objective of the Research

The main aims of this research work are to achieve the following objectives:

i. To propose and develop an approach to derive support tests as an alternative to the existing complete and sufficient tests proposed by previous researchers with the intention to increase the number of local checking for a given update operation.

ii. To propose a model to rank the several alternative tests so that that the most suitable test that reduces the amount of data transferred, the amount of data accessed, and the number of sites involved and finally achieves local checking processing, is selected to be evaluated.

1.4 Research Scope

The scope of this research work is outlined in the following points:

- The relational data model is used in this research work, because it is the most dominant model among the conventional models in terms of installed system (Feras, 2006; Thomas and Carolyn, 2005).
- The types of integrity constraints considered in this research work are *static integrity constraints* which include domain constraints, key constraints, referential integrity

constraints, and general semantic constraints. These types of constraints are used frequently in database applications.

- The integrity constraints which are considered in this research are assumed to be specified in Conjunctive Normal Form (CNF).
- The research work focuses on three types of integrity tests, namely: complete, sufficient, and support.

1.5 Methodology of Research

The research methodology of this research work can be summarized as follows:

- Literature Review –
    - The problem of constraint simplification is extensively deliberate and several available approaches have been reviewed. These include constraints simplification approaches proposed for centralized, distributed, parallel, and mobile databases.
    - The types of integrity tests in database systems have been reviewed.
- Design the Proposed Approach for Simplifying the Integrity Constraints –
    - The phases of constraints simplification process are proposed.
    - An approach to simplify the integrity constraints is proposed.
- Design a Model for Ranking and Classifying the Integrity Tests –
    - The phases for ranking and selecting integrity tests are proposed.
    - An approach for ranking and classifying integrity tests is proposed.
- Implementation –

- The phases of the proposed constraints simplification approach are implemented. In addition, the phases of the proposed ranking and classifying integrity tests approach are implemented.

- Evaluation –
  - Evaluate the performance of the proposed approaches for integrity constraint checking in distributed database system with respect to the amount of data transferred across the network, the cost of checking the constraints, and the percentage of local processing in checking integrity constraints.

1.7 Organization of the Thesis

This thesis is organized as follows. Chapter 2 presents the fundamental concepts of database systems. Further, the relational model which represents the data model in this thesis is illustrated. Moreover, the integrity constraints concepts are explained. The types of integrity constraints in the database systems are presented as well.

Chapter 3 presents brief discussion for types of semantic integrity subsystem in the database system. An overview for methods of checking integrity constraints, and the properties of integrity tests are reported. It also reviews the relevant works proposed by previous researchers of constraints simplification for centralized, distributed, parallel, and mobile databases systems.

Chapter 4 comprised the integrity constraints simplification process. The phases of the proposed simplification approach are explained. The algorithm of generating integrity

tests with several examples for a given integrity constraint and an update template are presented and discussed.

An approach for integrity tests ranking and classification is introduced in Chapter 5. This approach has four phases are introduced and discussed. The proposed model and the algorithm for ranking integrity tests are illustrated and explained with various scenarios.

Chapter 6 describes the components of the proposed framework of integrity constraints checking in distributed database system. The phases of the proposed constraints simplification and the ranking and classifying integrity tests are embedded in these components. In this chapter the results of the performance analysis are discussed and recorded. Three types of experiments are introduced with respect to three different criteria, namely: the amount of data transferred across the network, the cost of checking the constraints, and the percentage of local processing when various combinations of integrity tests are selected.

Conclusions and future works as guidelines for further research that can be added to the work drawn from the thesis are summarized in Chapter 7.

CHAPTER 2

DATABASE AND INTEGRITY

2.1 Introduction

This chapter presents an overview of the distributed database systems, the relational data model of a database, and the integrity constraints. This chapter is organized as follows: In Section 2.2 we present the concepts of database system which include the types of database system. Section 2.3 introduces the basic concepts of relational model in the database. Section 2.4 presents the concepts related to integrity constraints which include types of integrity constraints and their specification. In Section 2.5, an example database is presented. The database is used throughout this thesis.

2.2 Distributed Database System

In (Elmasri and Navathe, 2007), Fonkam (1993) and Ibrahim (1998), three broad classes of database system are identified. They are:

i. The conventional centralized databases where the information bases are logically centralized and physically centralized.

ii. The distributed databases where the information bases are logically centralized and physically decentralized.

iii. In the province of federated, information bases are logically decentralized and physically centralized or decentralized.